



Diplomado Data Science + Big Data

Jesús Ramos
TW: @xuxoramos
FB: /xuxoramos
Email: jesus@datank.ai

Roadmap

1. Fundacional / Tronco Común

- a. Agile Data Science
- b. R
- c. Python
- d. Data Wrangling
- e. Intro a Big Data

2. Track Data Science

- a. Aprendizaje Automático
- b. Aprendizaje Automático a Gran Escala

3. Track Data Engineering

- a. Hadoop
- b. Spark
- c. Stack SMACK

¿Cómo trabajaremos?

1. Sesiones + Capstone
2. Capstones en equipos de 4 personas, balanceados entre capacidad y acceso a datos
3. Bitbucket como control de versiones
4. ClouderaVM + Tools
5. Sesiones de Capstone con comunicación directa con mentor
6. Cada capstone tendrá un owner de negocio y un mentor
7. La comunicación será por Slack

¿Cómo pido ayuda?

1. José Incera (Coordinador Diplomado) - cuando quieras escalar issues, mándale correo.
2. Jesús Ramos (Subcoordinador Diplomado) - cuando tengas broncas con algún material, profe, o infraestructura del curso fuera de la de Citi.
3. María Asunción (Directora Extensión) - cuando ninguno de los de arriba te pueda solucionar las cosas.
4. Aurora Peral (Coordinadora Extensión) - cuando tengas problemas con cuestiones como asistencia o material.

Quién soy?

1. ISC02 (ITESM).
2. Financial Econometrics (UNottingham + UWashington)
3. Graduado de la Data Science Specialization de Coursera + JHU.
4. Consultado con +6 firmas para levantar capacidades analíticas (BMV, Indeval, GBM, ConCrédito, Propiedades.com, Nestlé, Telefónica, GNP, etc).
5. Cofundador de @TheDataPub, la comunidad de Ciencia de Datos más grande de México.
6. COO en Datank.ai.
7. Anti-buzzwords, anti-hype: Data Gestapo.
8. Gamer los sábados. Foodie los domingos.

Parte I

La economía de Datos

La economía de datos

Las fuerzas digitales

La velocidad industrial

La velocidad digital

Escenarios de transición

Fuerzas de la era digital



Fuerzas de la era digital

**Mundo
hiperconectado**



Fuerzas de la era digital

Hiperpersonalización

Mundo
hiperconectado



Fuerzas de la era digital

Hiperpersonalización

**Mundo
hiperconectado**



**Internet como medio
unificador**

Fuerzas de la era digital

Hiperpersonalización

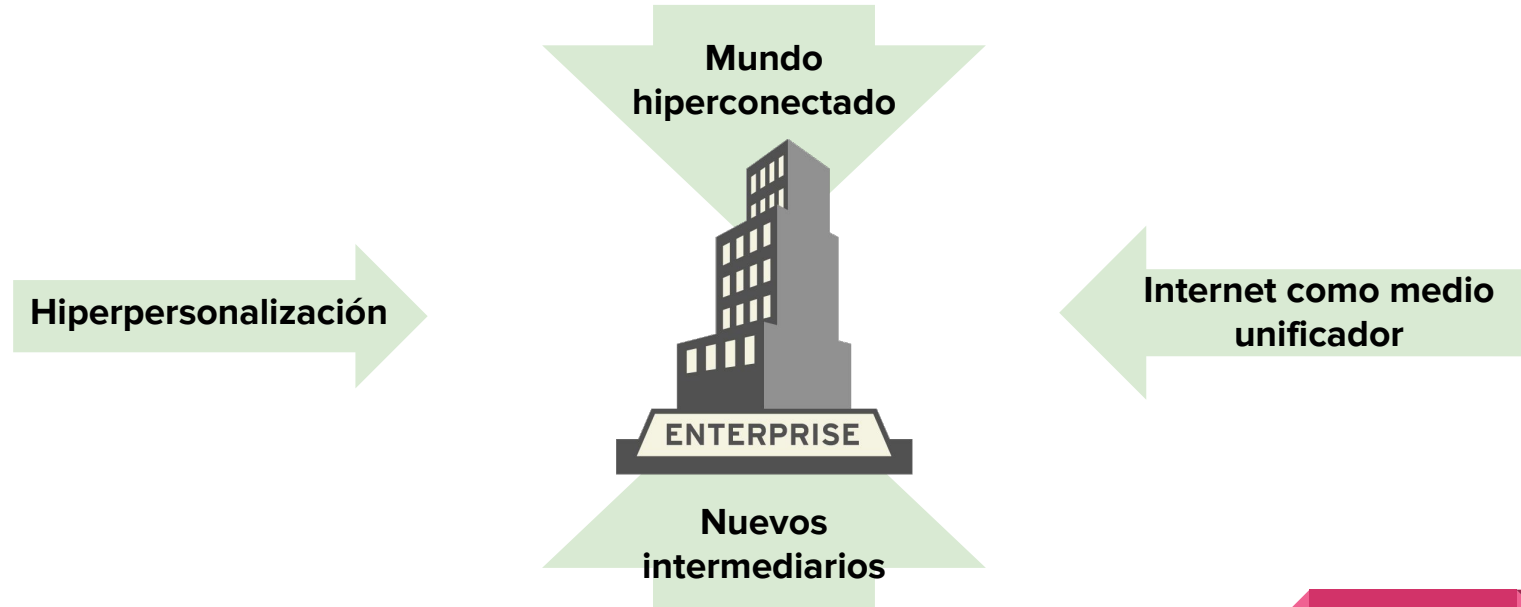
**Mundo
hiperconectado**



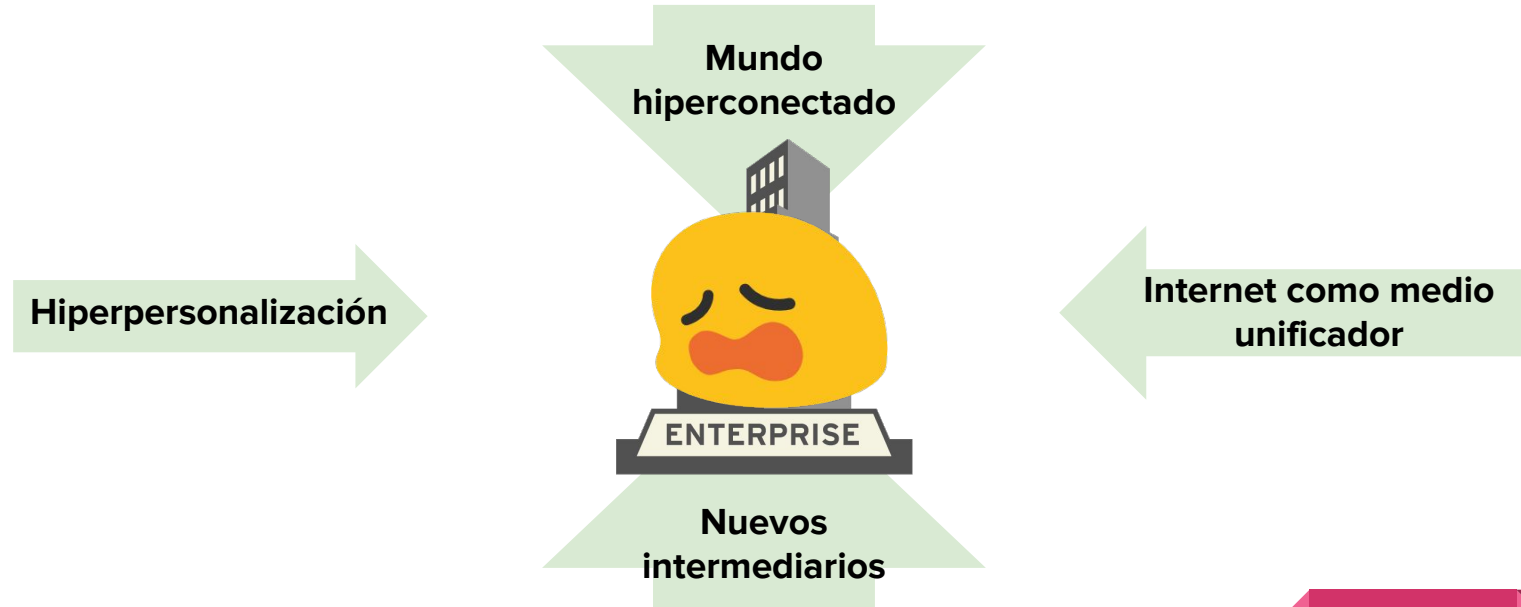
**Nuevos
intermediarios**


**Internet como medio
unificador**

Fuerzas de la era digital



Fuerzas de la era digital





Cómo fortalecemos a la org
para enfrentar estos retos?



Data!

Solo de la buena.

**The
Economist**

MAY 6TH-12TH 2017

Crunch time in France

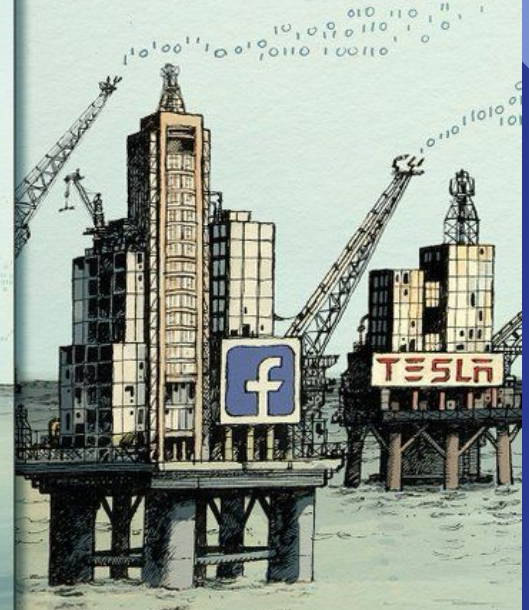
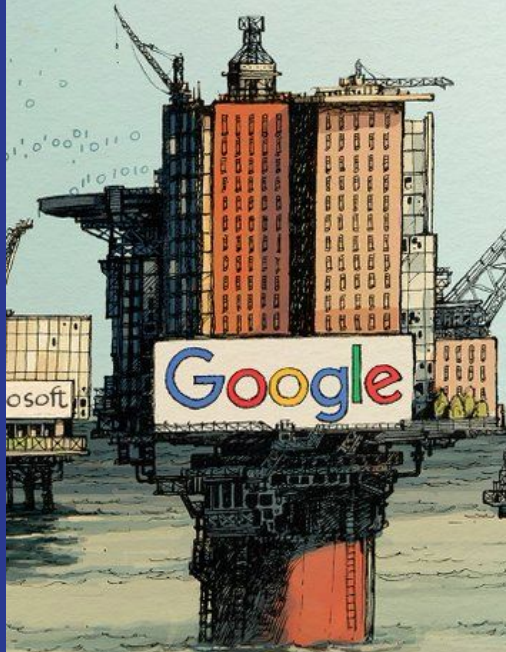
Ten years on: banking after the crisis

South Korea's unfinished revolution

Biology, but without the cells

The world's most valuable resource

Data and the new rules
of competition



Cuánto vale esto?

Cuánto vale esto?



\$19mmdd

Cuánto vale esto?



\$19mdd



\$2.5mdd

Cuánto vale esto?



\$19mdd



\$2.5mdd



\$990mdd

Cuánto vale esto?



\$19mmdd



\$2.5mdd



\$990mdd

Ca\$h Flow?

Cuánto vale esto?



\$19mdd



\$2.5mdd



\$990mdd

Ca\$h ~~X~~ Flow?

Cuánto vale esto?



\$19mdd



\$2.5mdd



\$990mdd

DATA*

* <https://techcrunch.com/2015/10/13/whats-the-value-of-your-data/>

Stakeholders tecnológicos

Reguladores

Autoridades

Cámaras



Clientes

Proveedores

Business Partners

Las 2 velocidades del IT

Velocidad 1



Velocidad 2

Las 2 velocidades del IT

Velocidad 1

Solidez y
confiabilidad

Estabilidad y
compliance



Velocidad 2

Las 2 velocidades del IT

Velocidad 1

Solidez y
confiabilidad

Estabilidad y
compliance



Velocidad 2

Agilidad y
experimentación

Fluidez



Las 2 velocidades del IT

Indeed, the ability to offer new products on a timely basis has become an important competitive factor; this might require **weekly software releases** for an e-commerce platform.

That kind of speed can only be achieved with an inherently error-prone software-development approach of **testing, failing, learning, adapting, and iterating rapidly**.

It's hard to imagine that experimental approach applied to legacy systems. Nor would it be appropriate, because the **demand for perfection is far higher in key back-end legacy systems**.

Las 2 velocidades del IT

	<u>Industrial IT (Speed 1)</u>	<u>Digital IT (Speed 2)</u>
Tipo de solución	Madura, Legacy	Hecha a la medida
Metodología	Waterfall + Estándares	Agiles
Objetivo	Eficiencia Operativa	Ventaja competitiva
Atributo clave	Estabilidad	Velocidad
Propósito esencial	Dar confianza	Enganchar y deleitar
Camino que cubre	2/3	1/3
Rol de IT	"Keep lights on"	Socio de innovación
Quién lo dirige?	CTO, CIO	CMO, Chief Digital

Caso de Implementación Virtuosa



1. Modelo traído desde España
2. CTO queda responsable de sistemas backend (Velocidad 1)
 - a. SPEI
 - b. Cámaras de compensación
 - c. Conexión con INDEVAL y Bolsa Mexicana
 - d. Reportería a CNBV
3. Chief Innovation responsable de customer-facing tech (Velocidad 2)
 - a. Apps
 - b. Open APIs
 - c. Ecosistema emprendedor
4. APIs son solo front-end contracts de sistemas Legacy y soportados (no diseñados) por equipos del CTO
5. **Creación de servicios con ensamble de APIs crea sinergias entre CTO y Chief Innovation.**

Caso de Implementación Incompleta



1. 2 organizaciones paralelas
 - a. GBM Corporativo
 - b. GBM Digital
2. CTO reporta a GBM Corporativo
 - a. Responsable de sistemas legacy
 - b. Responsable de desarrollo
 - c. Responsable de infraestructura
3. GBM Digital depende de CTO de Corporativo
 - a. Apps se diseñan en Digital y se implementan en Corporativo
 - b. Deficiente arquitectura de sistemas no permite separación ni experimentación
 - c. **APIs y Apps se desarrollan con mismas metodologías y herramientas que sistemas y motores de misión crítica.**
 - d. Lenta respuesta a algunos proyectos de innovación.

Caso de Implementación Deficiente



1. Junta directiva vetusta: "así lo hemos hecho por 50 años".
2. 17 años de retraso en iniciativas digitales
3. Imposición de modelo organizacional desde NY
4. Estrés organizacional por reacomodo de estructura
5. Estructura en espejo sin distinguir funciones
 - a. Citi NY nombra Chief Data Officer del lado de IT
 - b. Citibanamex nombra Chief Data Officer del lado de Analytics
 - c. Repetición de funciones
 - d. Confusión y desánimo en los equipos
6. Resistencia interna al cambio de herramientas como SAS y HP Vertica.
 - a. Falta de capacidades internas
 - b. Erosión del talent pool disponible
7. **Elevadísimo costo reputacional y organizacional para innovar.**

Parte II

Arquitectura de Datos de 2 velocidades

- DWH y la velocidad 1
- Intro al Data Lake
- El Data Lake y la velocidad 2
- Integrando Data Lake + DWH

Las 2 velocidades del IT

Velocidad 1

Solidez y
confiabilidad

Estabilidad y
compliance



Velocidad 2

Agilidad y
experimentación

Fluidez



Las 2 velocidades del IT

	<u>Industrial IT (Speed 1)</u>	<u>Digital IT (Speed 2)</u>
Tipo de solución	Madura, Legacy	Hecha a la medida
Metodología	Waterfall + Estándares	Agiles
Objetivo	Eficiencia Operativa	Ventaja competitiva
Atributo clave	Estabilidad	Velocidad
Propósito esencial	Dar confianza	Enganchar y deleitar
Camino que cubre	2/3	1/3
Rol de IT	"Keep lights on"	Socio de innovación
Quién lo dirige?	CTO, CIO	CMO, Chief Digital

Ejercicio: Necesidades de IT de 2 velocidades

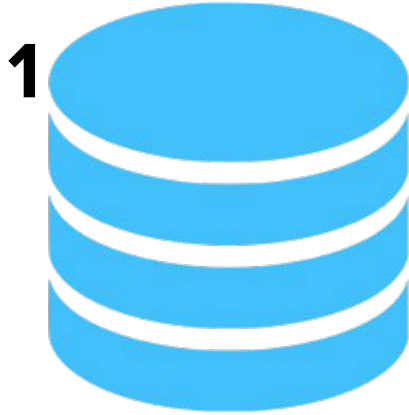
Velocidad 1

Velocidad 2

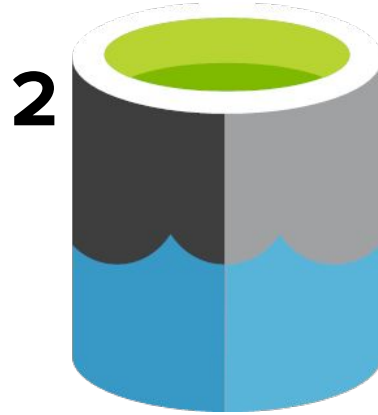


Ecobici

Artefactos de IT para 2 velocidades

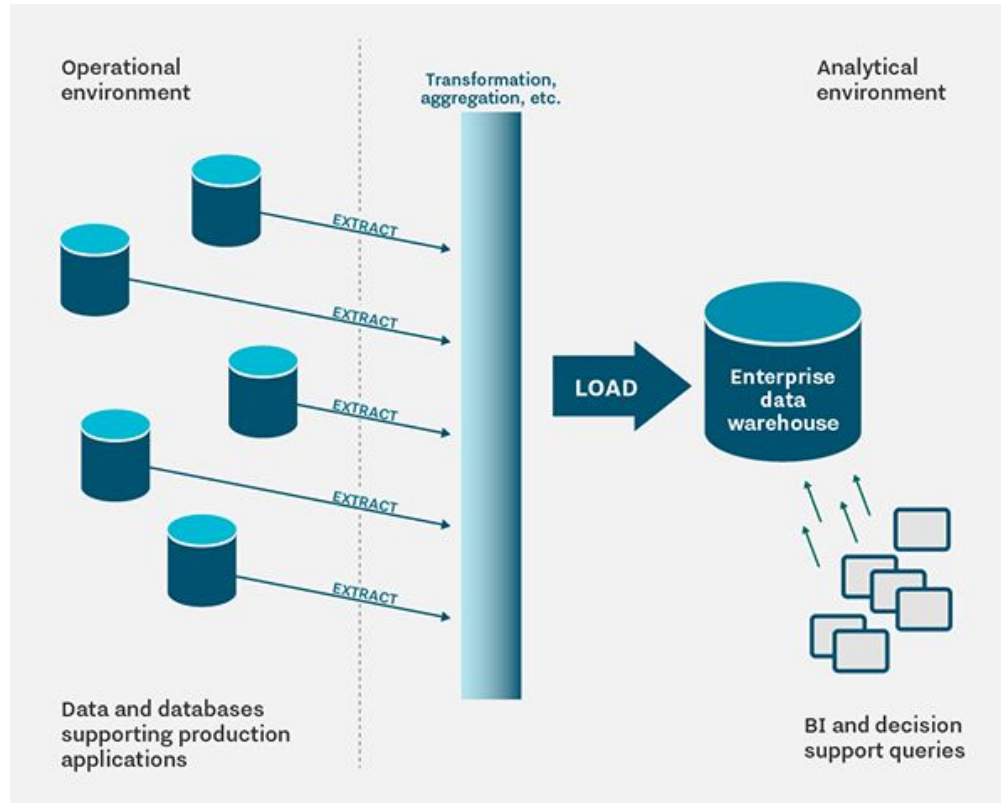


Enterprise Data Warehouse

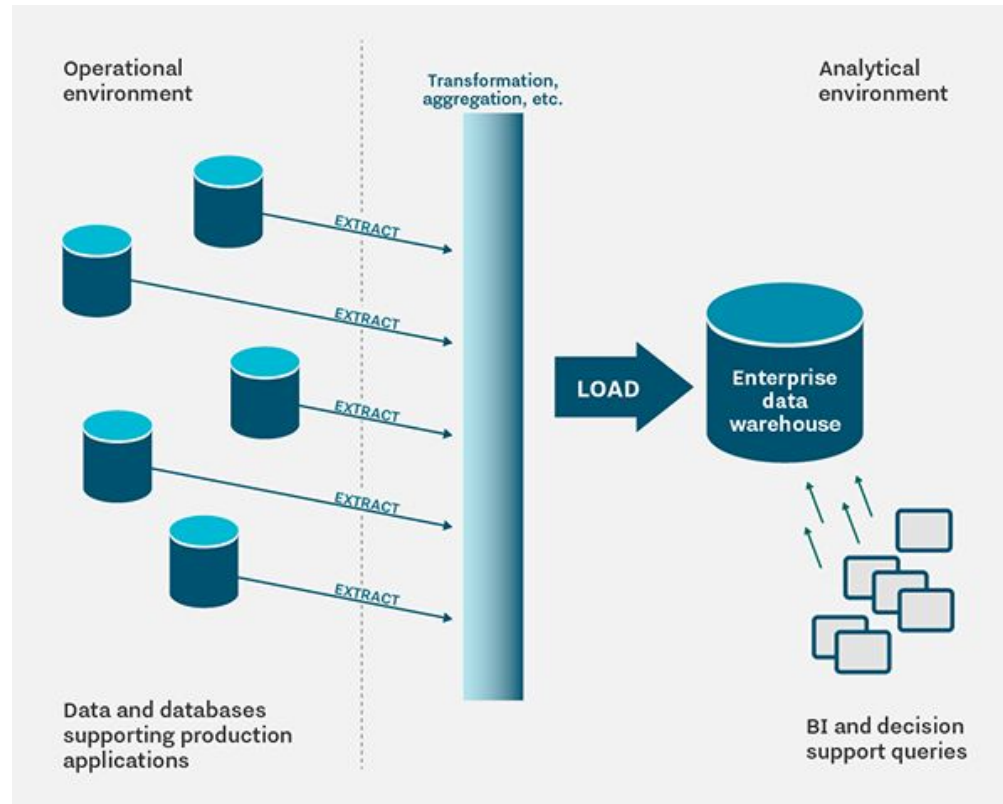


Enterprise Data Lake

Definición de Data Warehouse



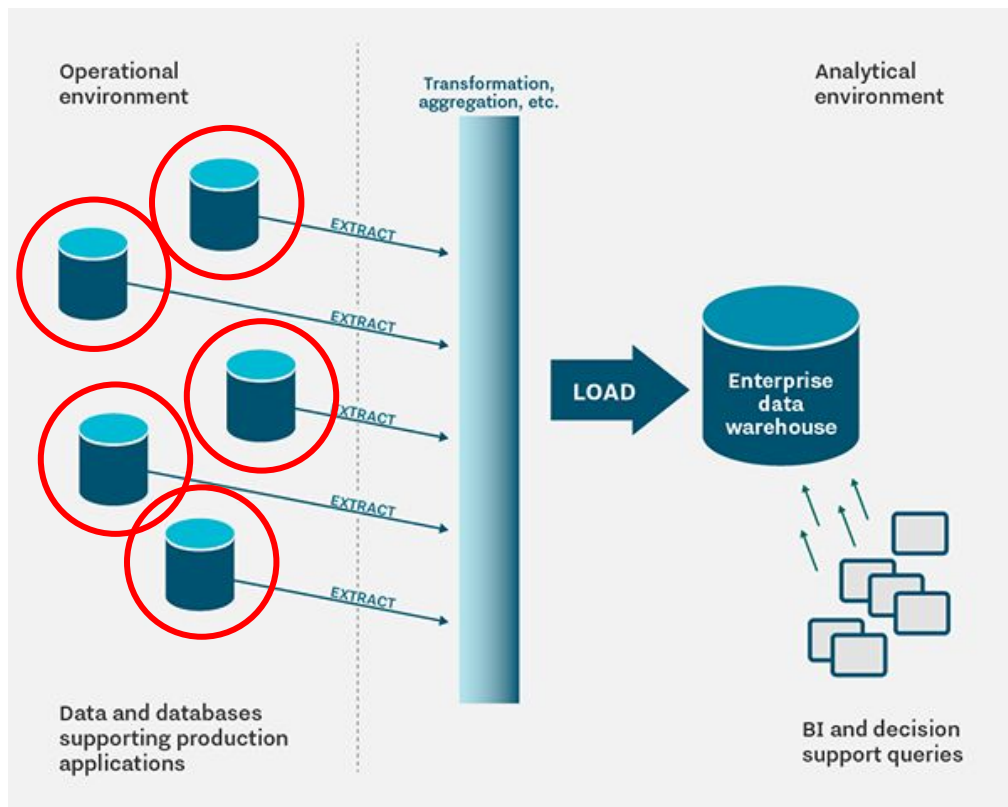
¿Dónde está la velocidad 1?



¿Dónde está la velocidad 1?

5 Esquemas Diferentes

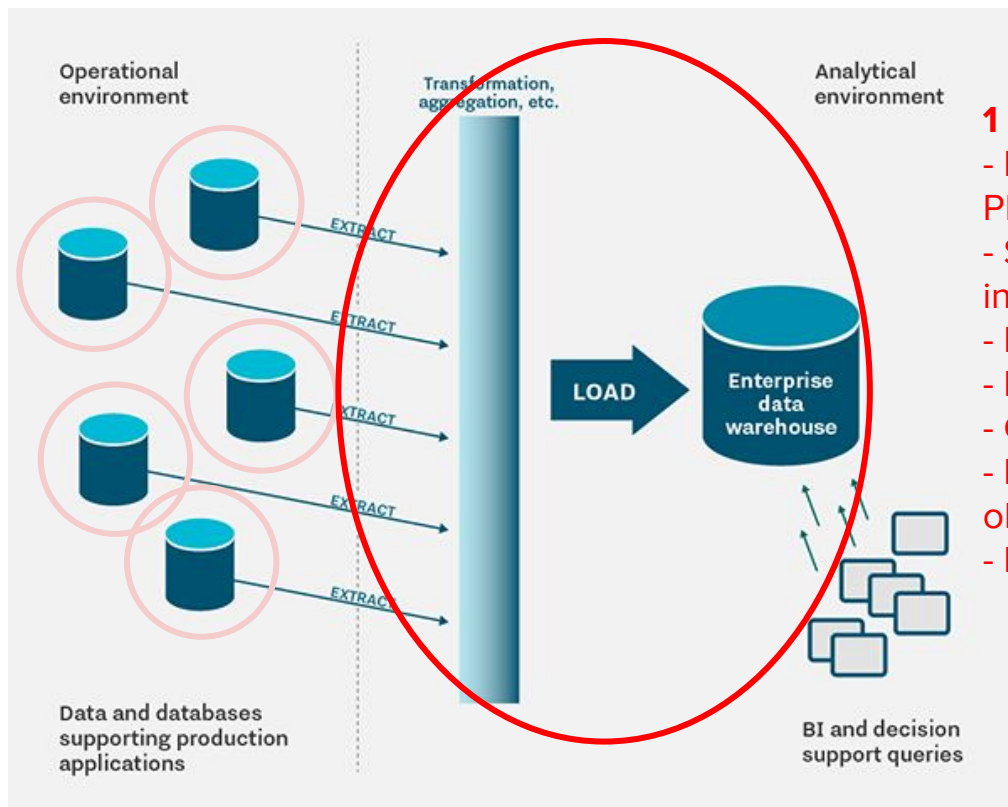
- Sin catálogos compartidos
- De propósito específico
- Diferentes niveles de normalización
- Mismos datos, poco governance
- Registros duplicados
- Columnas repetidas



¿Dónde está la velocidad 1?

5 Esquemas Diferentes

- Sin catálogos compartidos
- De propósito específico
- Diferentes niveles de normalización
- Mismos datos, poco governance
- Registros duplicados
- Columnas repetidas



1 Esquema Unificador

- Establecido PREVIAMENTE
- Sin esquema no hay integración
- De propósito general
- Desnormalización
- Governance obligatorio
- Desduplicación obligatoria
- Perfilamiento obligatorio

¿Dónde está la velocidad 1?

5 Esquemas Diferentes

- Sin catálogos compartidos
- De propósito específico
- Diferentes niveles de normalización
- Mismos datos, poco governance
- Registros duplicados
- Columnas repetidas



1 Esquema Unificador

- Establecido PREVIAMENTE
- Sin esquema no hay integración
- De propósito general
- Desnormalización
- Governance obligatorio
- Desduplicación obligatoria
- Perfilamiento obligatorio

¿Dónde está la velocidad 1?

5 Esquemas Diferentes

- Sin catálogos compartidos
- De propósito específico
- Diferentes niveles de normalización
- Mismos datos, poco governance
- Registros duplicados
- Columnas repetidas



1 Esquema Unificador

- Establecido PREVIAMENTE
- Sin esquema no hay integración
- De propósito general
- Desnormalización
- Governance obligatorio
- Desduplicación obligatoria
- Perfilamiento obligatorio

El Data Lake

Diferencia con el Data Lake

WAREHOUSE

Cliente del Lake

Estructurado / Con Esquema

Esquema fijo previo a Lectura

Caro para grandes volúmenes

Maduro (copiado de las RDBMS)

BI / Business Analysts

Menor (config fija)

SQL-compliant

Estrictos (reportería normativa y operativa)

Lugar en la org

Tipos de Datos

Lectura

Costo Almacenamiento

Seguridad

Usuarios

Agilidad

Flexibilidad de herramientas

SLAs

Fuente del DWH

Cualquiera

Esquema volátil al leer

Bajo costo en el tiempo

Emergente

Data Scientists / Engineers

Mayor (conf varias veces)

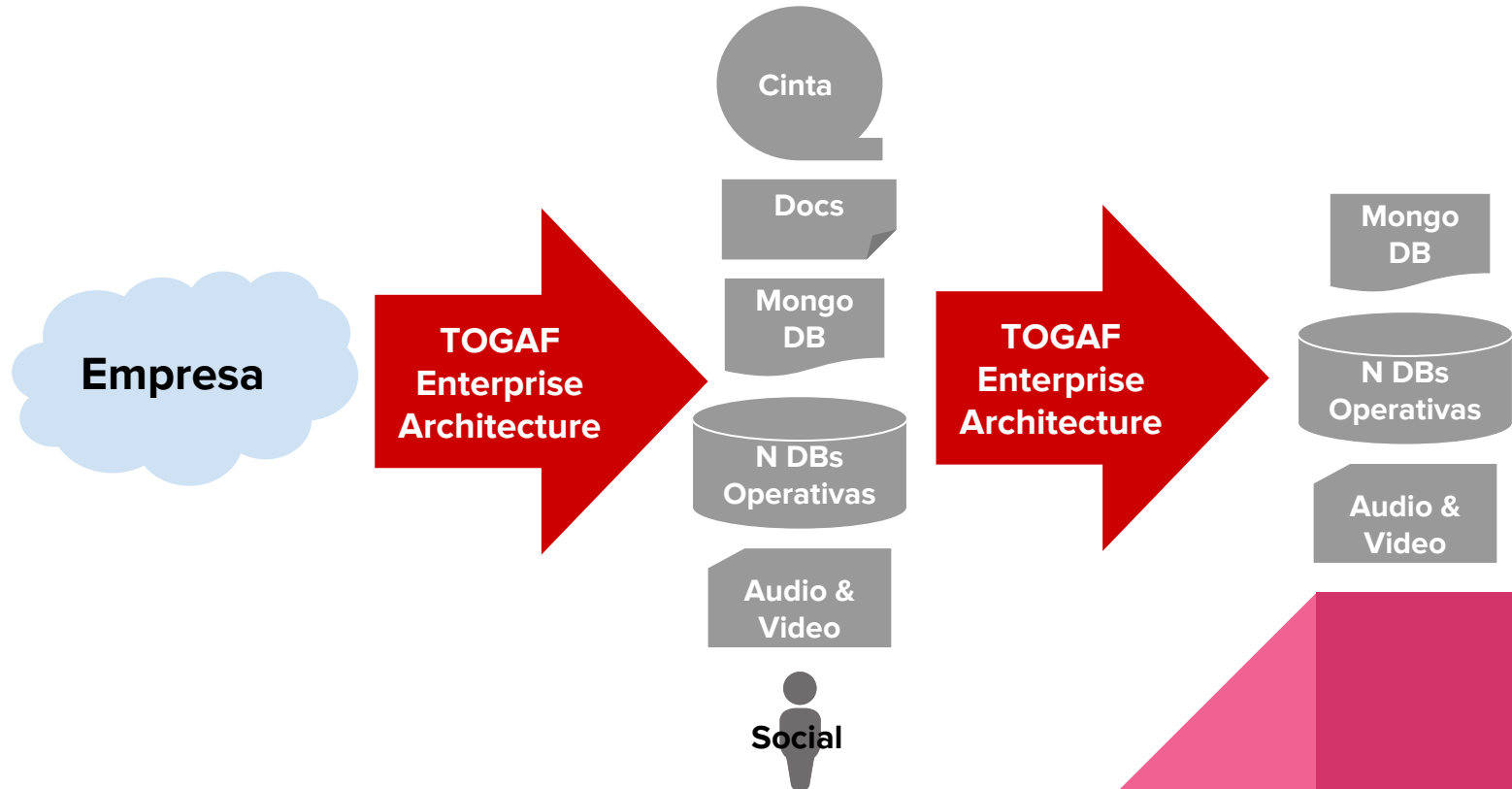
Varias (mayormente open source)

Laxos / Sin SLA

DATA LAKE

Arquitectura del Data Lake

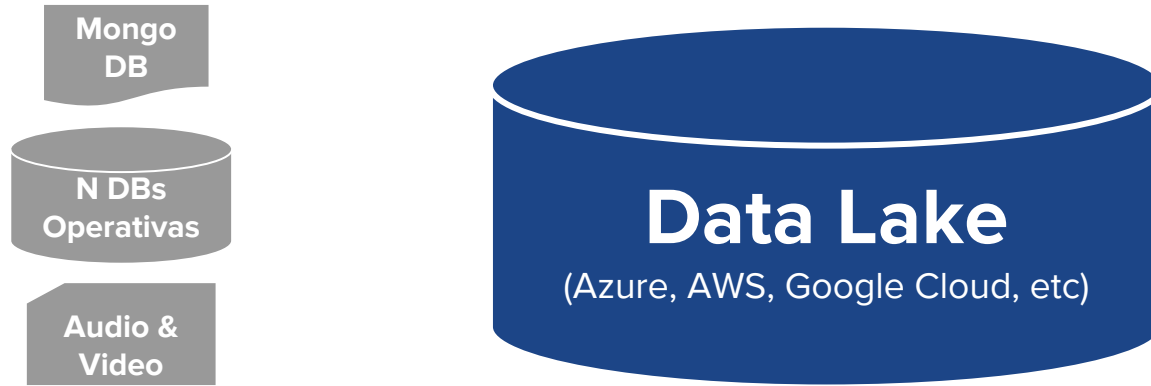
Identificación de Data Sources



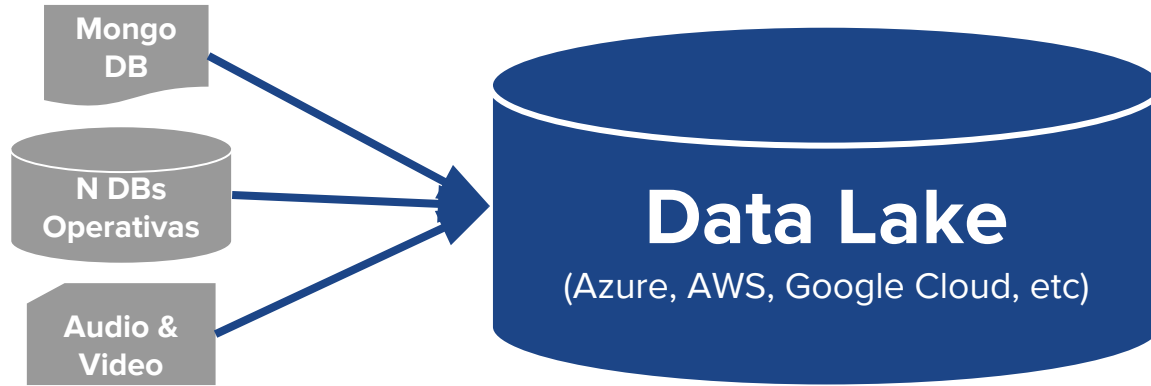
Aprovisionamiento



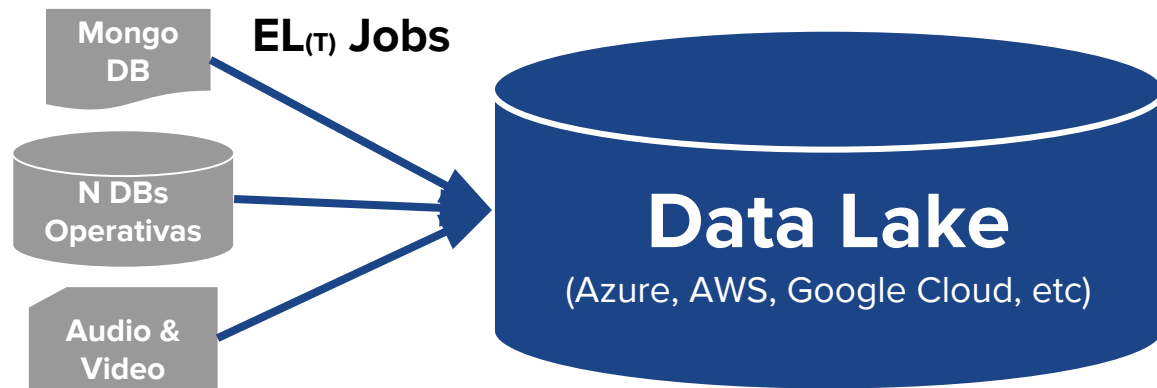
Creación de Extraction & Loading Jobs



Creación de Extraction & Loading Jobs



Creación de Extraction & Loading Jobs



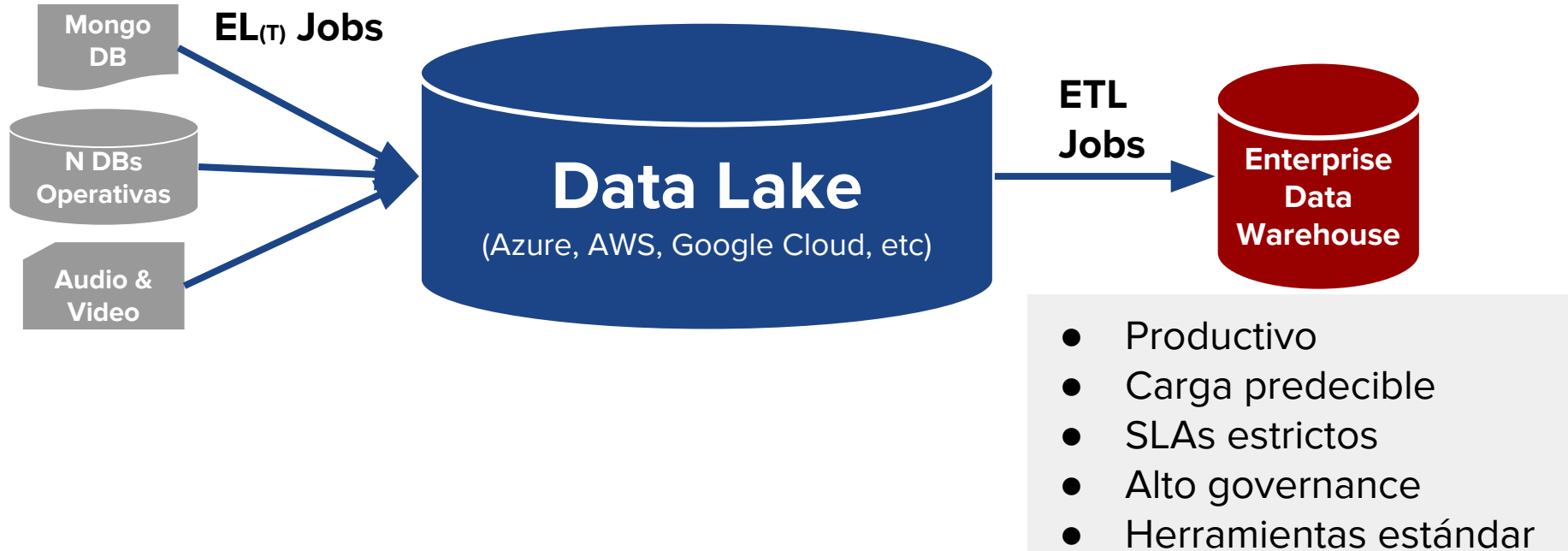
A diferencia de los ETLs, no hay transformación.

Las bases se cargan **AS-IS** y se versionan con **timestamp**.

Se guardan en **directorios** que reflejen la fuente.

Se etiqueta cada carga con **metadata** de origen.

El Regreso del DWH

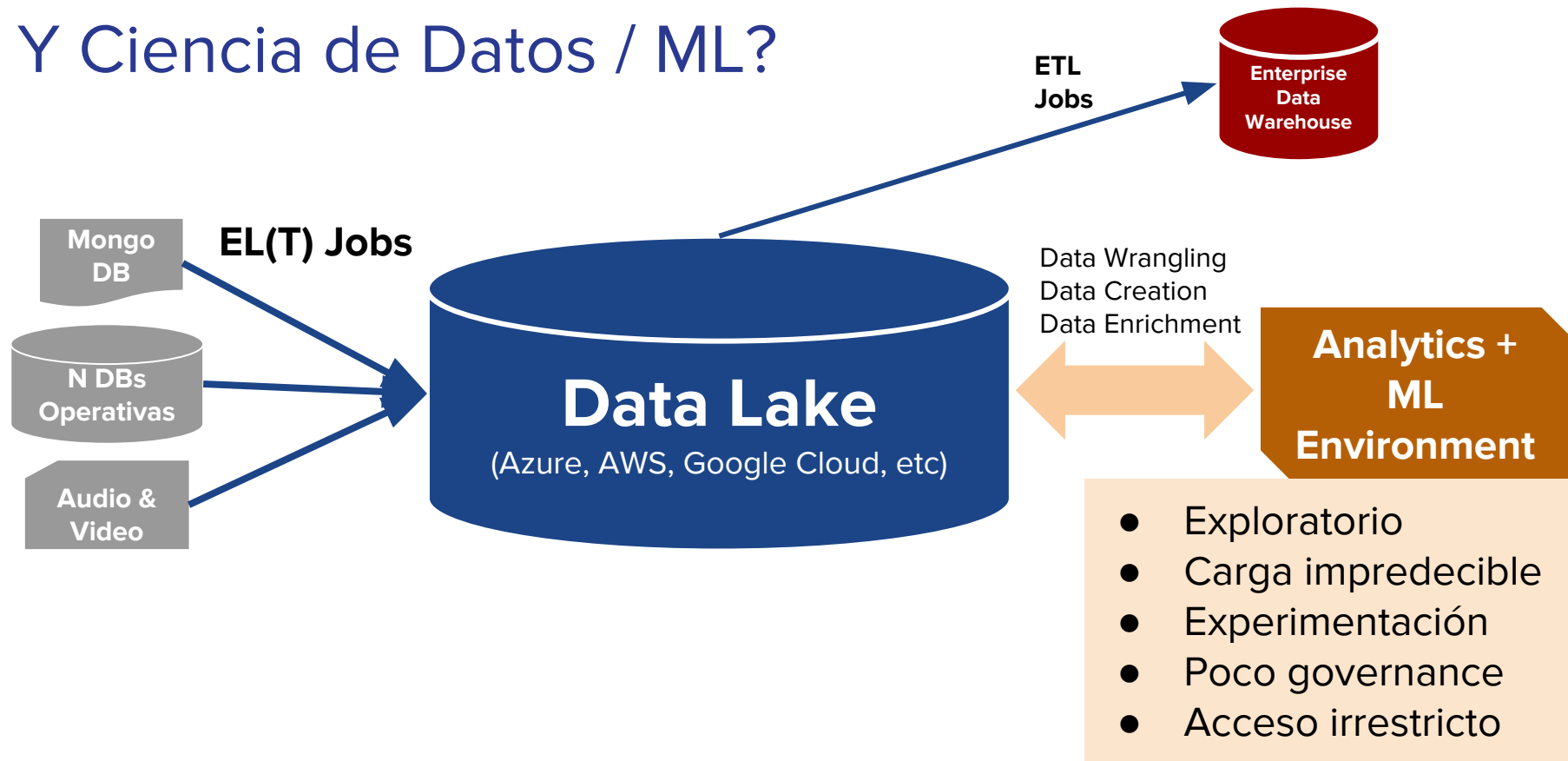


El Regreso del DWH

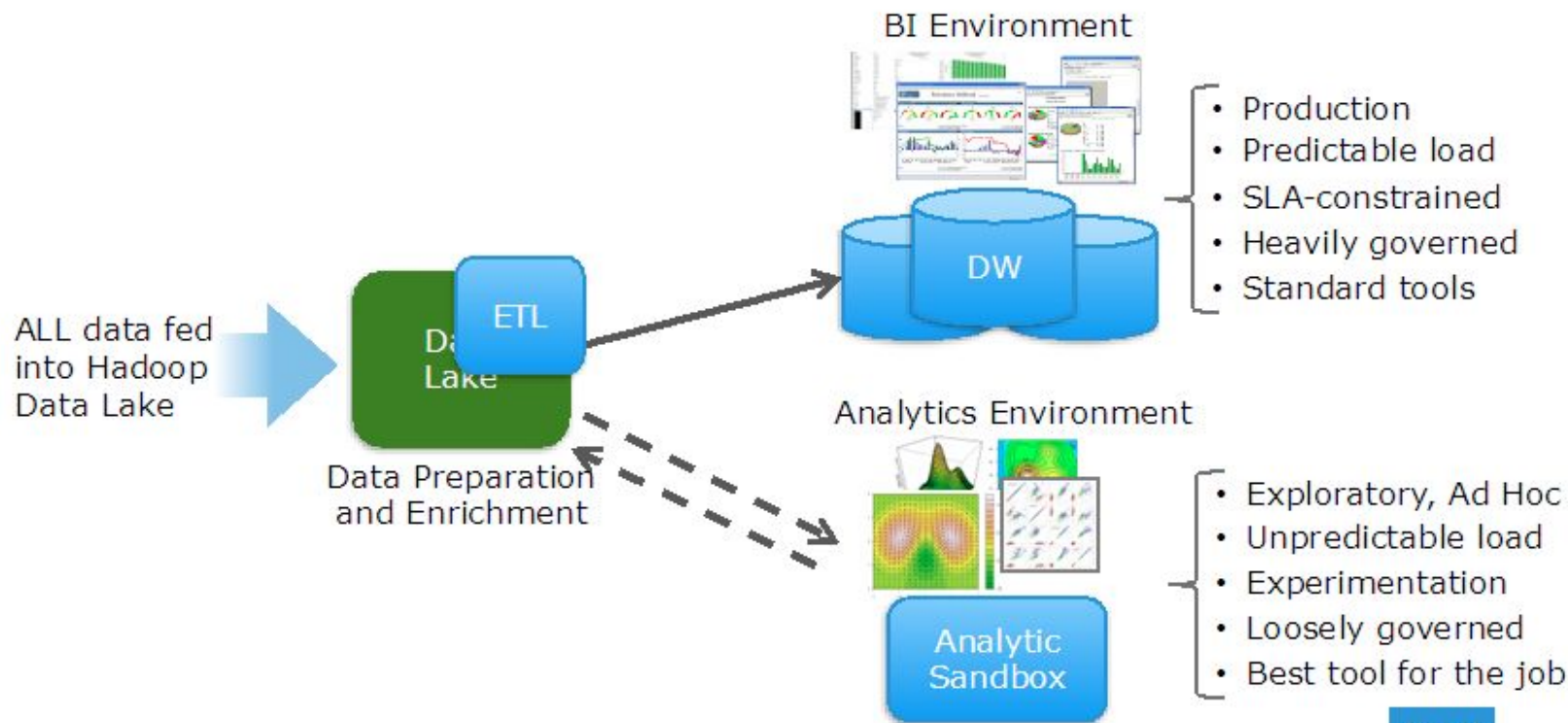


Imperativo hacer pruebas de no-impacto, dado que estamos cambiando las fuentes de las cuales se alimenta el DWH.

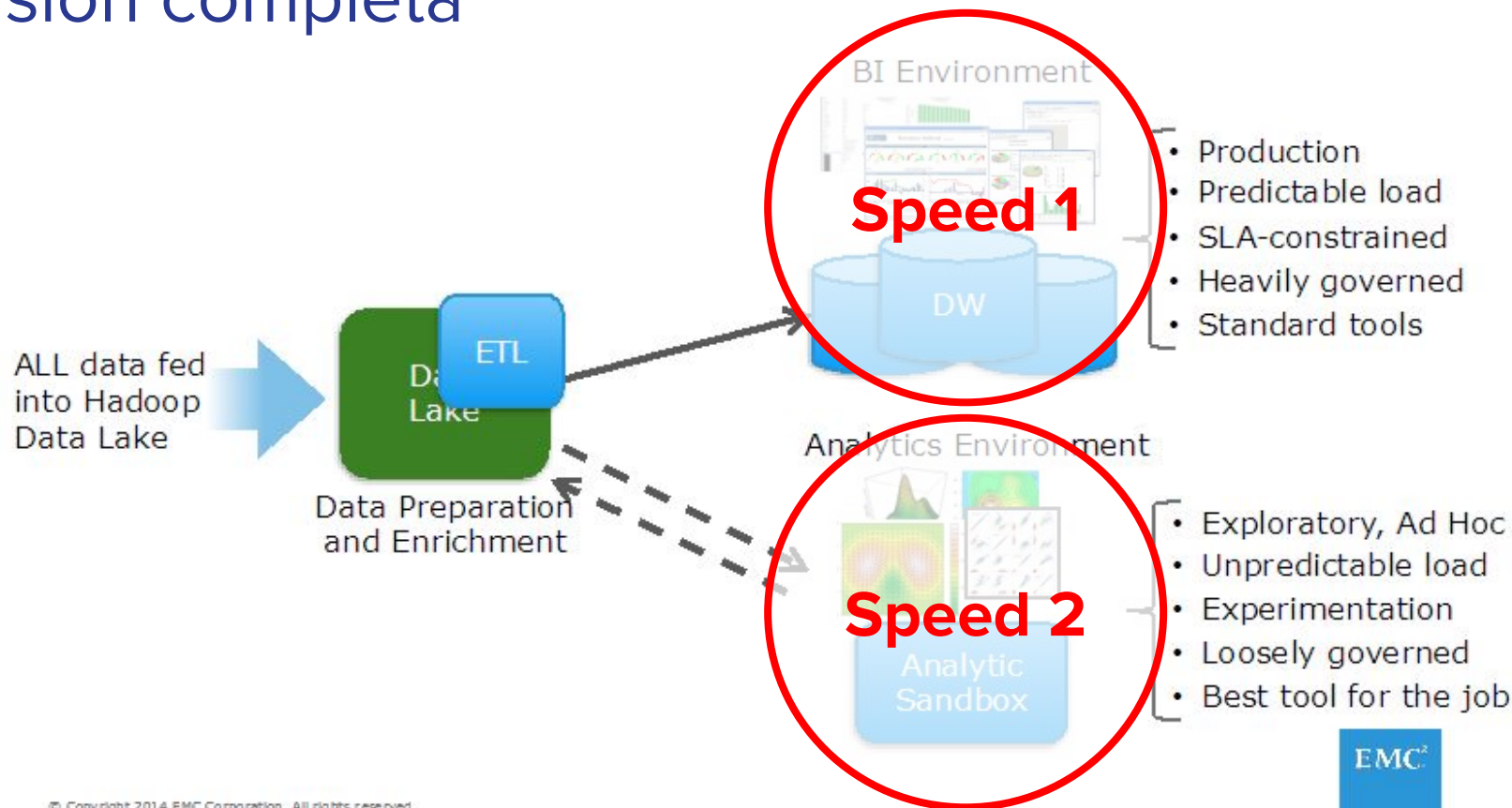
Y Ciencia de Datos / ML?



Visión completa



Visión completa



Parte II (cont)

Arquitectura de Datos de 2
velocidades

- Integrando Data Lake + DWH
- Para qué analizamos datos?
- Definición de Productos de Datos
- Metodología para desarrollo de productos de datos para velocidad 1

¿Para qué analizamos datos?

- 1. Logro de eficiencias operativas**
- 2. Desarrollo de nuevos productos**
- 3. Soporte a la toma de decisiones**

Diferencia entre BDs Transax y Analítica

TRANSAX

Operación diaria del negocio
Ejecutar tareas de negocio
Rápidas y en tiempo real
Simples y puntuales
Poco si no hay histórico
Muchas tablas normalizadas
Respaldar religiosamente

Fuente de datos
Propósito de los datos
Escrituras
Consultas
Espacio Requerido
Diseño de BD
Backup

La BD Transaccional
Decision, análisis, planeación
Tardadas y en batch
Complejas, con agregados
Mucho
Pocas sin normalización
Respaldos más relajados

ANALÍTICA

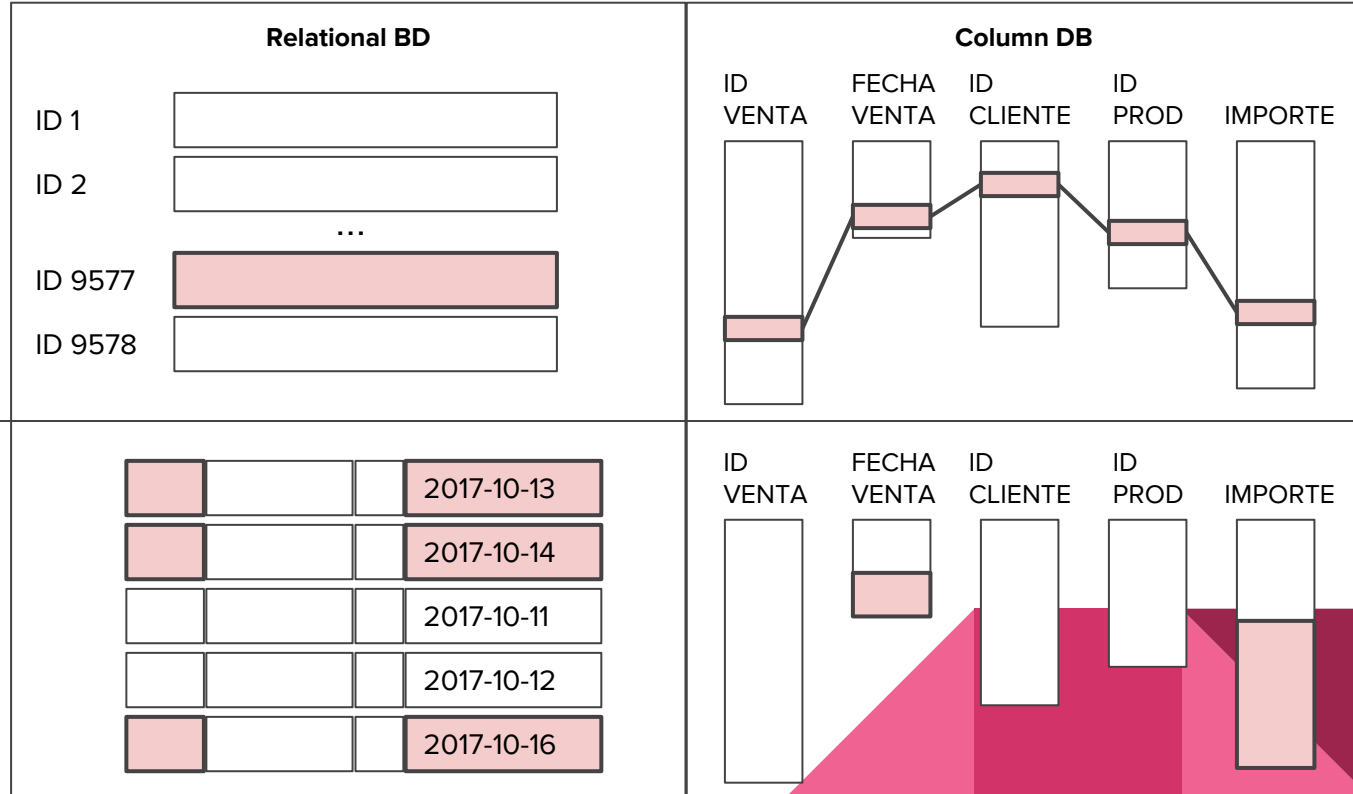
BDs Columnares vs Relacionales

```
SELECT * FROM VENTAS
WHERE ID_VENTA = 95771256
```

(típico query de sistemas transaccionales)

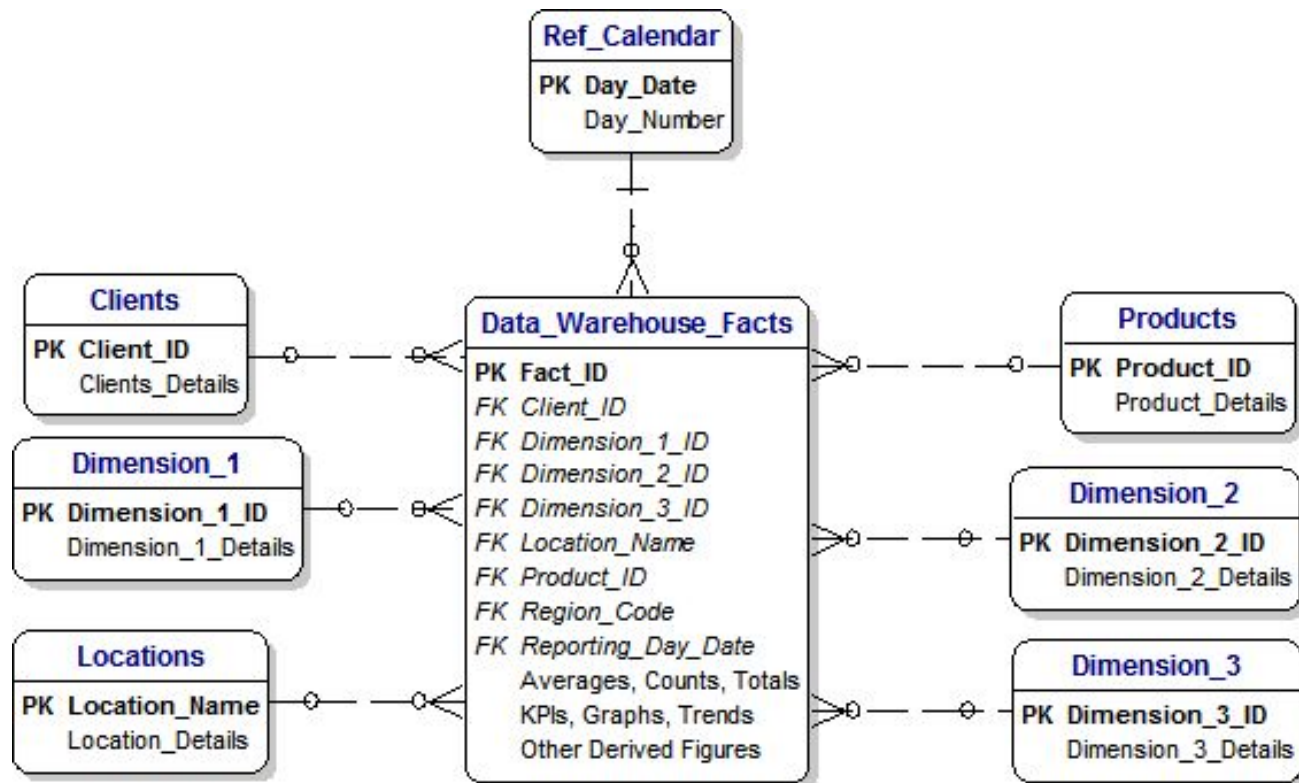
```
SELECT SUM(IMPORTE) FROM
VENTAS WHERE
FECHA_VENTA > 2017-10-12
```

(típico query analítico)



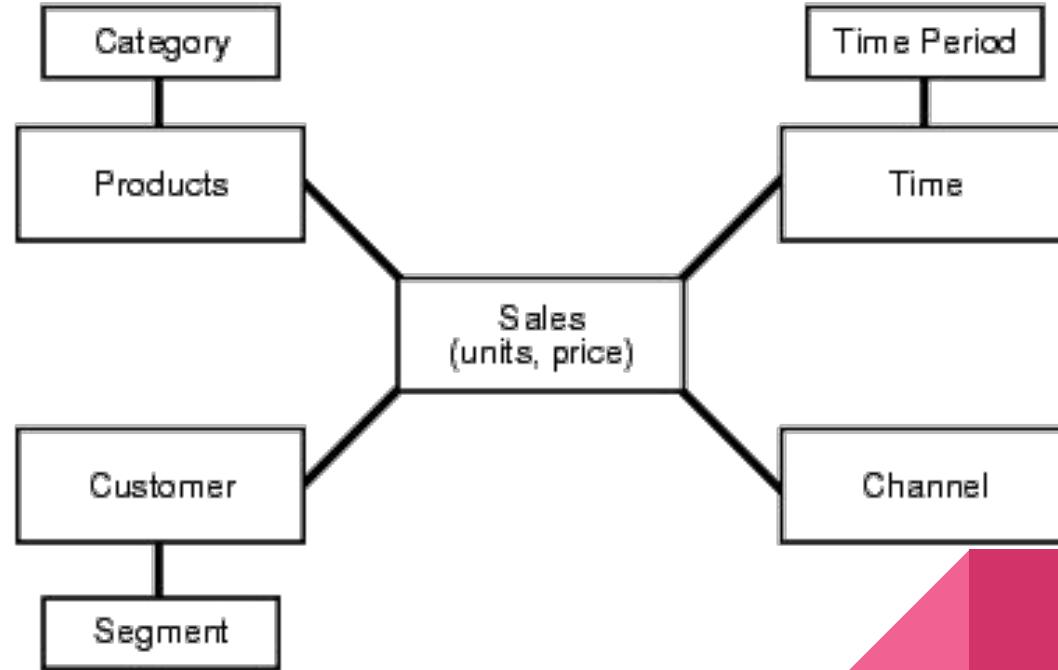
Modelos para construcción del DWH: Star

Star Schema



Modelos para construcción del DWH: Snowflake

Snowflake
Schema



Modelos para construcción del DWH: Big Table

Customer

CustID	FirstName	LastName	ContactInformation	ContactType
101	Elaine	Stevens	555-2653	Work
101	Elaine	Stevens	555-0057	Cell
102	Mary	Dittman	555-8816	Work
104	Drew	Lakeman	555-0949	Work
103	Skip	Stevenson	555-0650	Work
102	Mary	Dittman	555-8173	Fax
105	Eva	Plummer	Plummer@akcomms.com	Email
101	Elaine	Stevens	Stevens@akcomms.com	Email
101	Elaine	Stevens	555-5787	Fax
103	Skip	Stevenson	Stevenson@akcomms.com	Email
105	Eva	Plummer	555-5675	Work
102	Mary	Dittman	Dittman@akcomms.com	Email

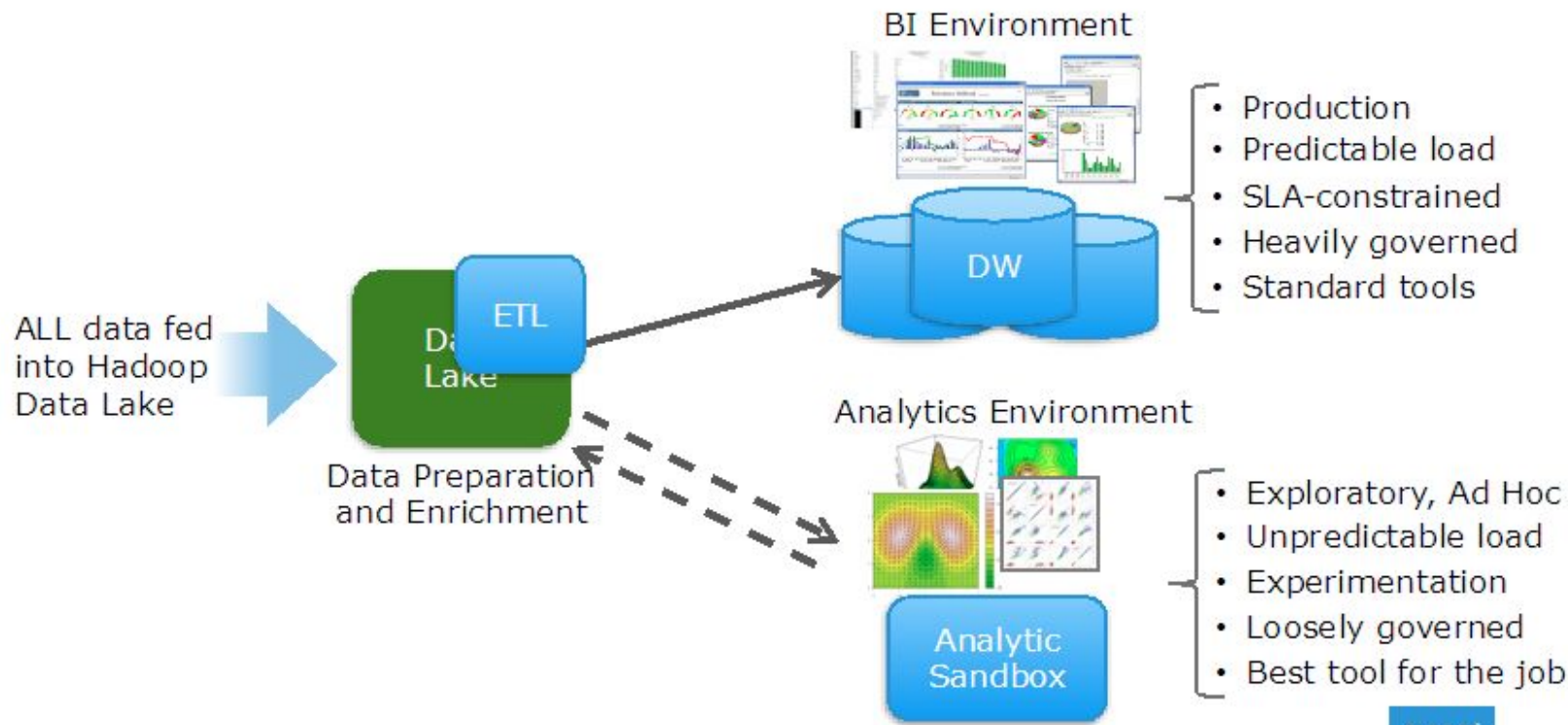
**Primary
Key**

**Atomic
Data**

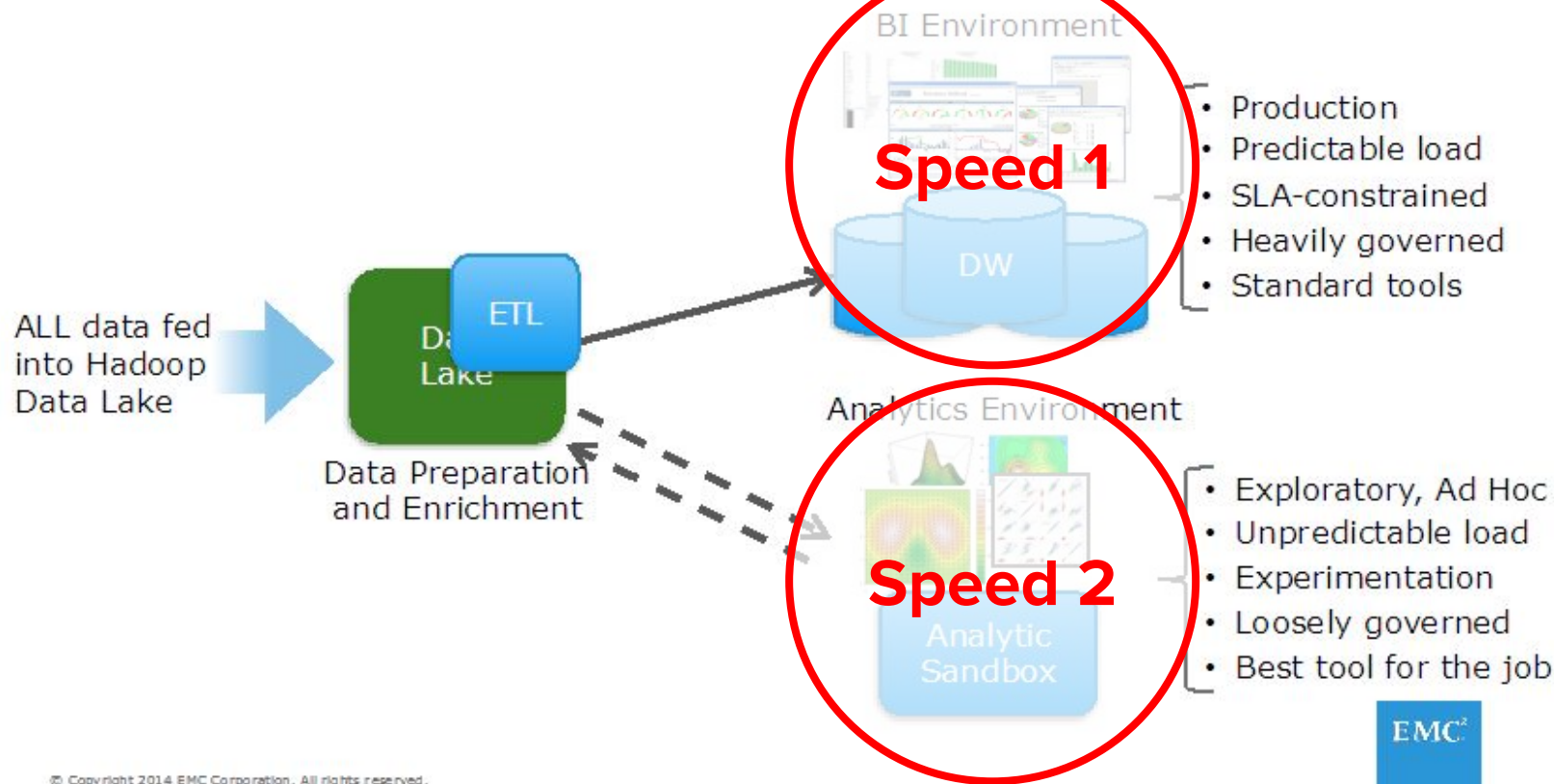
**Atomic
Data**

Big Table

Visión completa (2 velocidades, 1 artefacto)



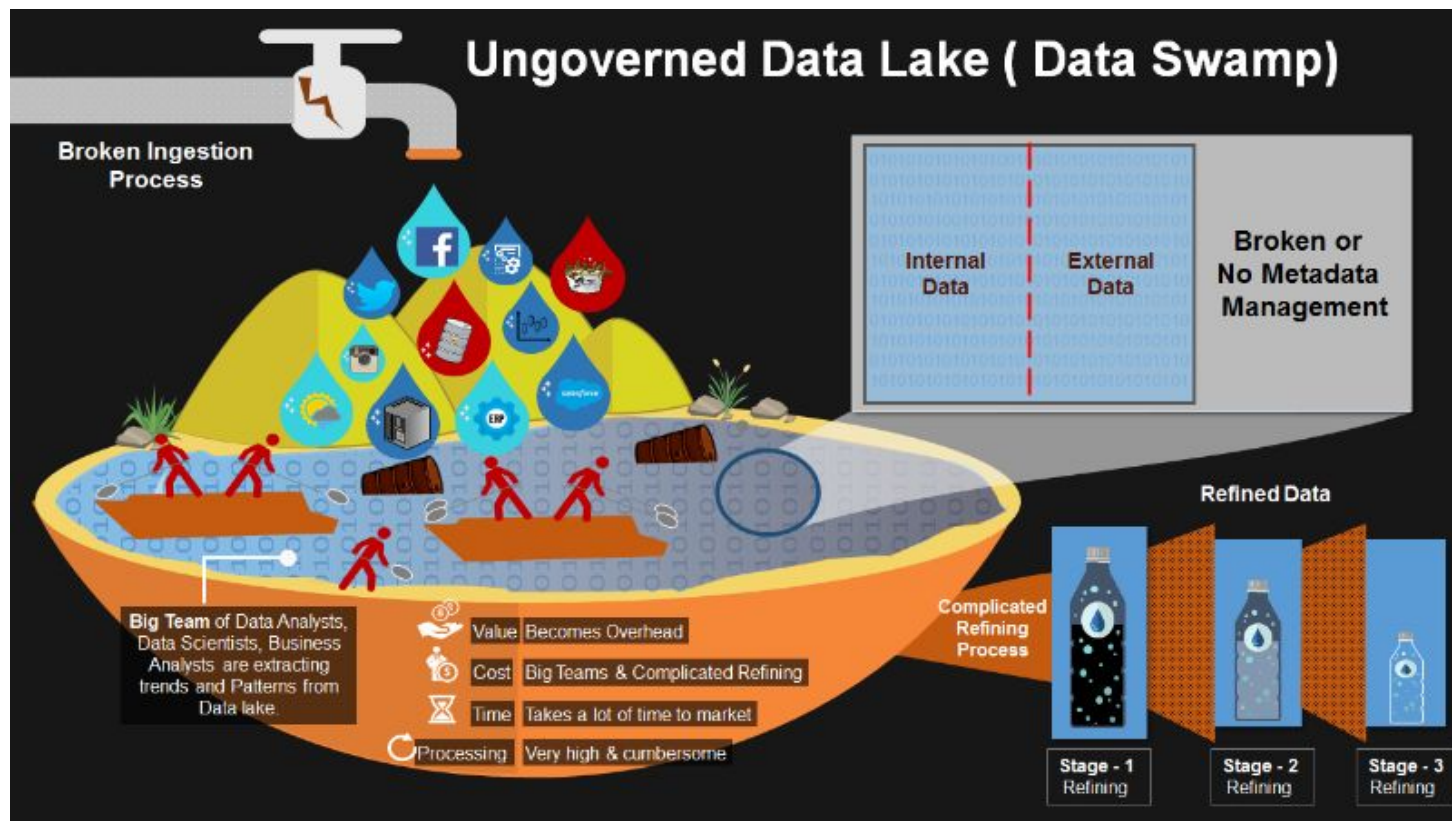
Visión completa (2 velocidades, 1 artefacto)



El Data Swamp

"Who pee'd in my Data Lake?"

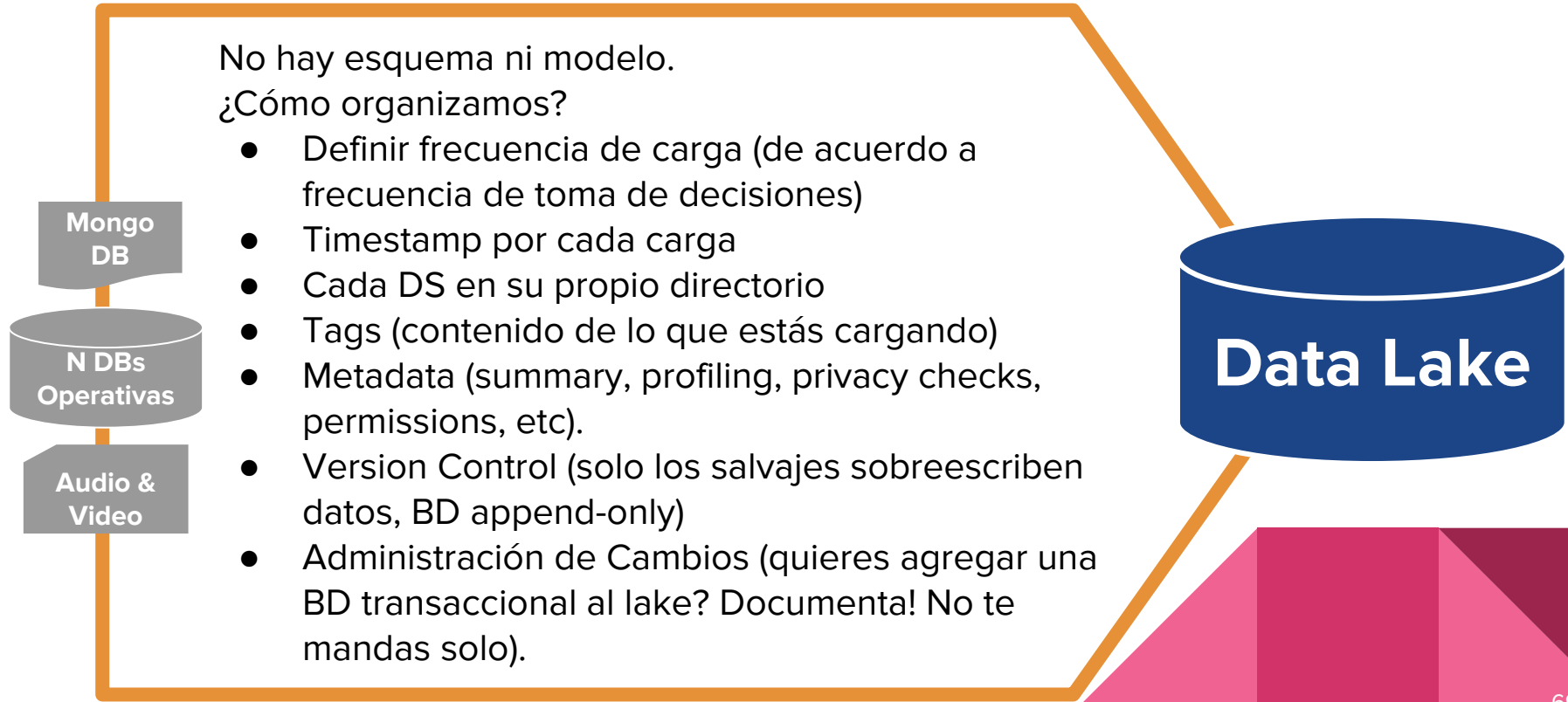
¿Qué convierte mi Data Lake en Xochimilco?





¿Cómo gobierno mi Lake?

Data Lake Gov'nance



Herramientas Open Source para DG

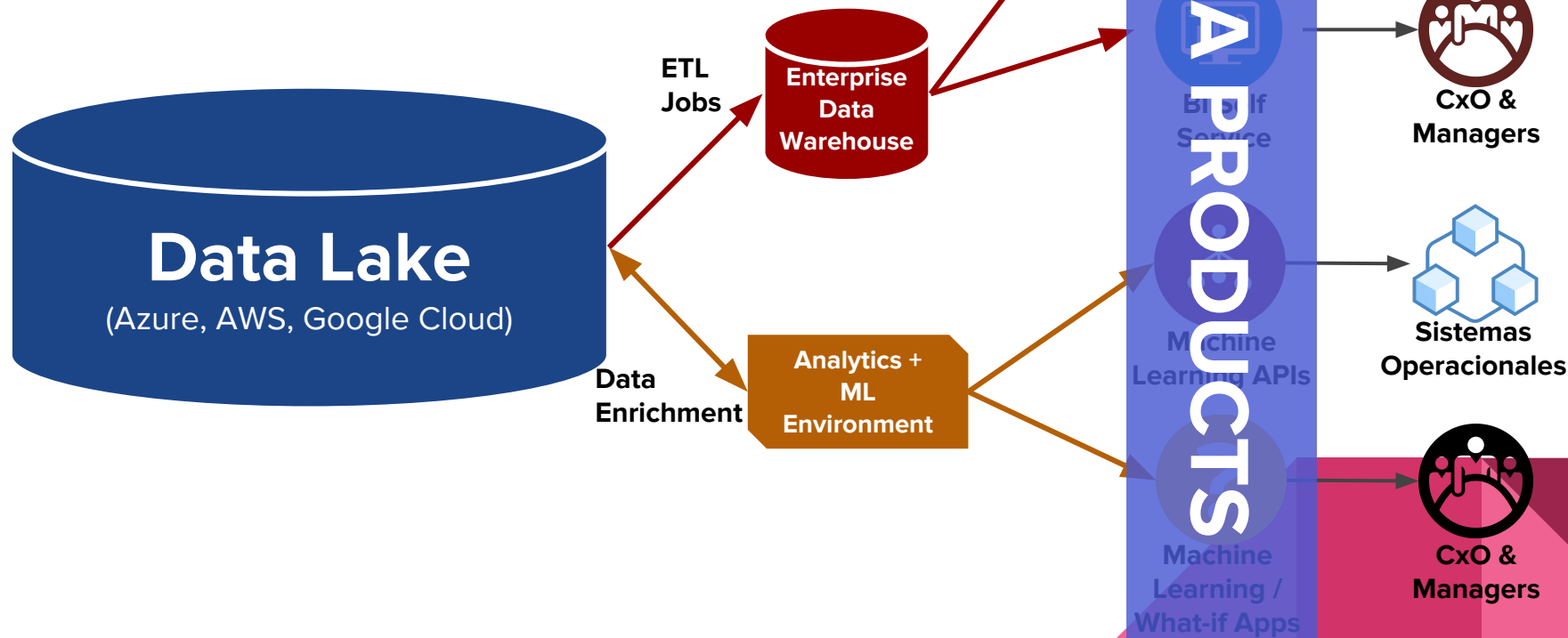


1. Solo cubiertos Privacy + Metadata + Lineage + Workflow
2. Seguridad en Hadoop: Apache Ranger
3. Workflow Management: Apache Airflow
4. Metadata Management & Data Lineage: Apache Atlas
 - a. Se integra con Ranger
 - b. Emanado de partes Open Source de Cloudera Navigator
5. Falta madurez
6. Áreas de Compliance/COBIT/ITIL no tienen su VoBo aún.

Ya tenemos los artefactos para 2 velocidades. Qué hacemos con ellos?

Productos de Datos!

¿Qué es un Data Product?



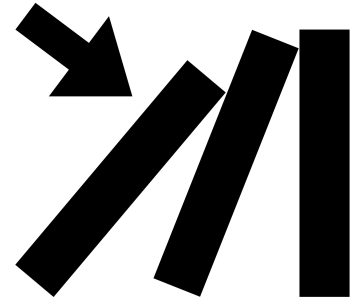
3 elementos para que un DP sea un DP



Observar la
realidad del
negocio
(datos consolidados y de
calidad)



Actuar sobre la
realidad
(que tiempo entre
conocimiento y decisión
tienda a 0)



Observar efecto
sistémico
(que los datos consolidados
se refresquen con el efecto
de la acción)

¿Qué NO es un producto de datos?

Reportes (usualmente de consultores chupasangre)

Informes (se guardan en libreros y acumulan polvo)

Exceles (lo que se hace ahí no regresa al DWH ni al Lake)

Power Points (se actualizan solos? puedo manipular los charts? Didn't think so...)

"**Modelo** es lo que el ingeniero implementa, no lo que el Científico de Datos crea."

- Dr. Adolfo de Unánue

Ejemplos de Productos de Datos (Speed 1)



- BMV tiene reuniones de consejo cada mes.
- Se discuten Mkt Cap, Índices y GDP de México y LATAM.
- Cada mes se invierten 320 hrs-hombre a razón de +\$100 USD/hora para producir un Power Point con esta info.
- La info no se discute porque está estática en los slides.
- BMV adquirió QlikView como BI de autoservicio interno.
- **Discusiones más ricas y decisiones más ágiles.**

Parte III

Introducción a Ciencia de Datos y Machine Learning

Qué es Ciencia de Datos?

Qué no es Ciencia de Datos?

Diferencias entre DS, DM, ML, AI, BI,.

Diferencia entre Desarrollo de SW y DS

Sesgo vs Varianza

Usos del ML

Tipos de ML

Y la AI?



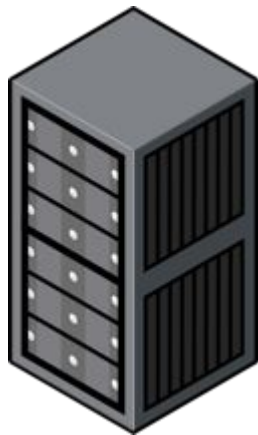
Entonces qué es
"big data"?

Definición "Big Data"

Una empresa que tiene:

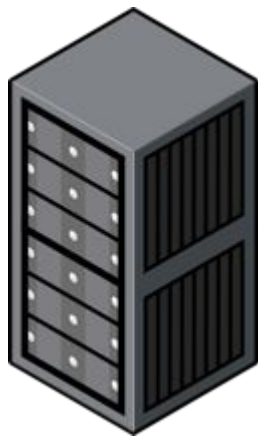
- Tablas (+120 columnas, 1M renglones, 6TB)
- Texto (2K páginas)
- Social (5K seguidores, 10 TW/posts x seguidor)
- Vídeo (20TB)
- Y puede/espera monetizar todo.
- Entonces si tiene Big Data.

Definición "Big Data"

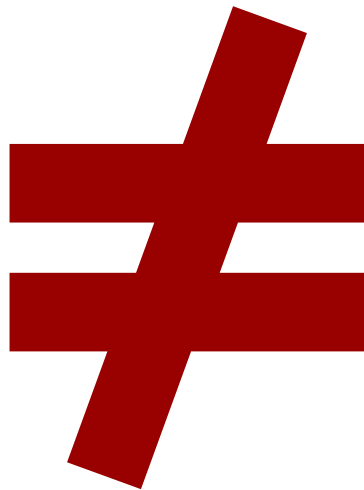


Servers, pipelines,
cloud infra, boxes
para procesar y
mover datos

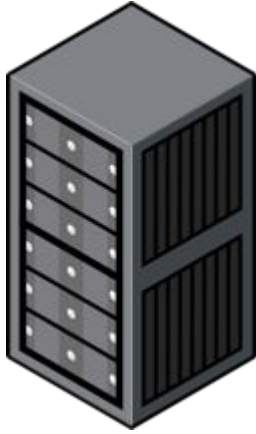
Definición "Big Data"



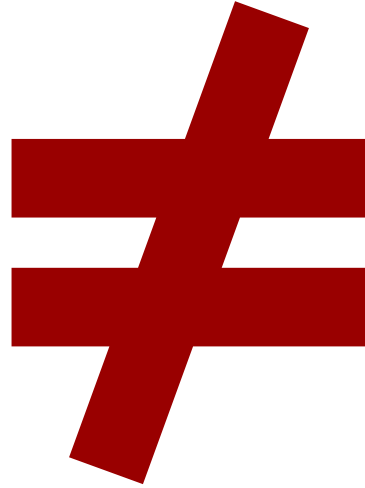
Servers, pipelines,
cloud infra, boxes
para procesar y
mover datos



Definición "Big Data"

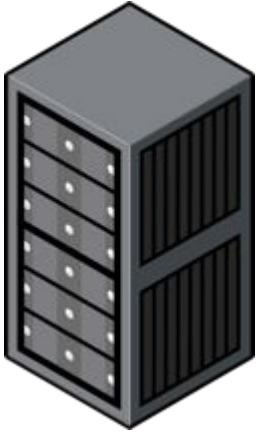


Servers, pipelines,
cloud infra, boxes
para procesar y
mover datos

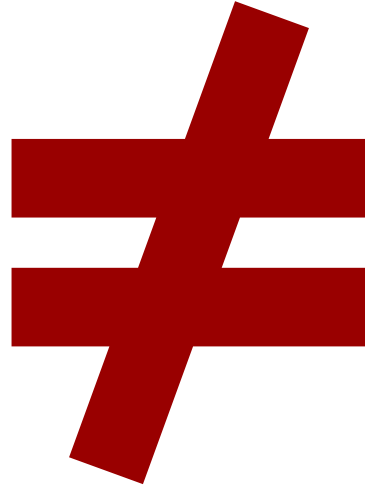


El conocimiento
que sacamos de
estos datos.

Definición "Big Data"



Big Data



Analytics



**"Big data needs Analytics.
Analytics doesn't need big
data"**

- Carla Gentry

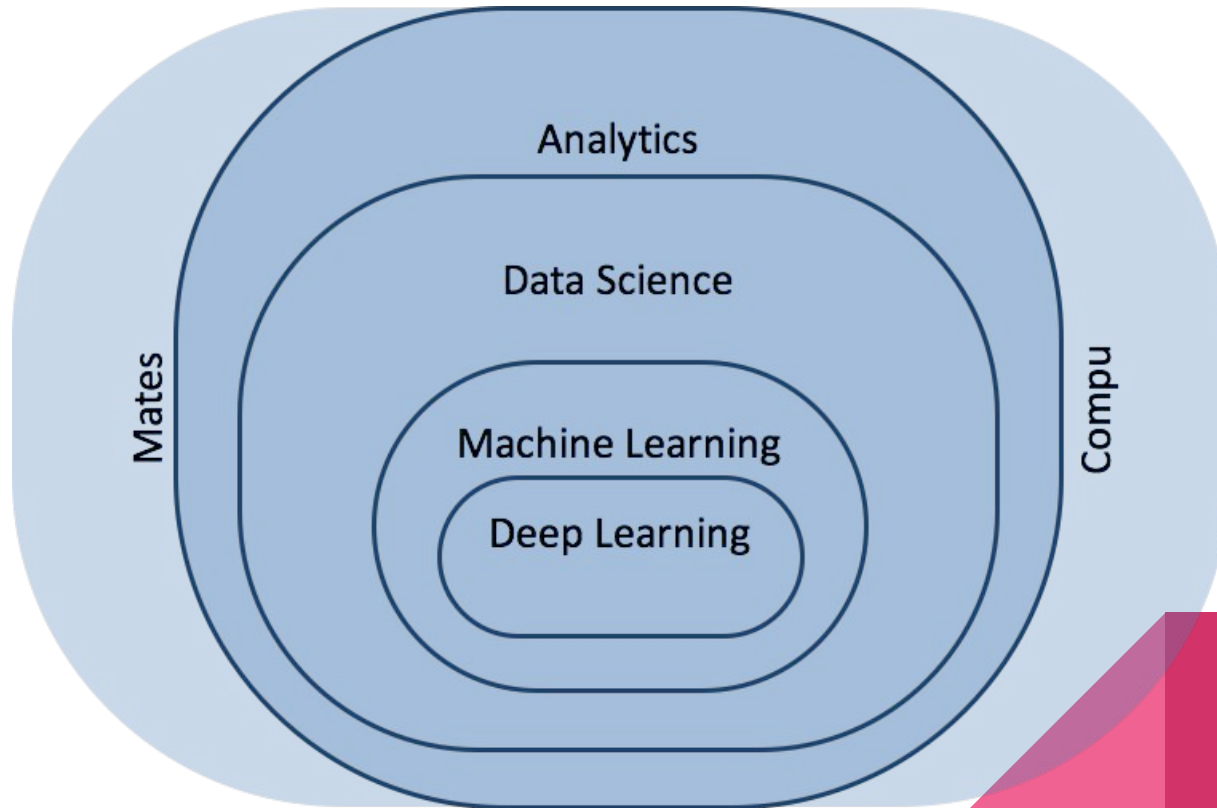
Ejemplo:

- Chico de 19 años de Actuaría UNAM baja datos de Ecobici de solo un par de meses y solo de sus las estaciones cerca de él.
- Baja sus datos a Google Sheets.
- Usa promedios móviles de num de viajes y capacidad de estación para decidir a cuál ir para garantizarse una bici en buen estado.
- No usó Big Data. Ni Deep Learning. Ni infraestructura cara. Ni nube.
- Usó Google Sheets, con un par de miles de registros, y promedios móviles, matemáticas de primeros semestres de licenciatura.
- Logró una eficiencia operativa para sí mismo.
- Mejoró su relación con un servicio público con costo 0.

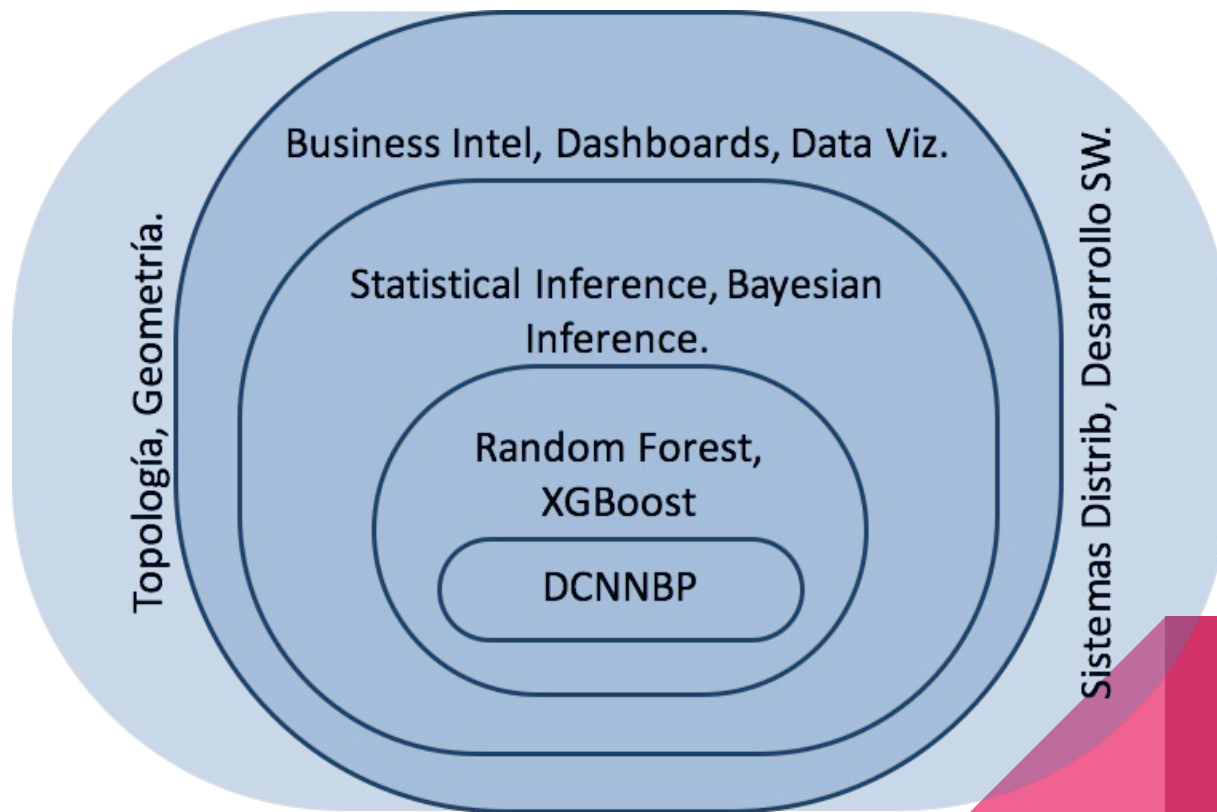


Qué es "Analytics"?

Qué es "Analytics"?

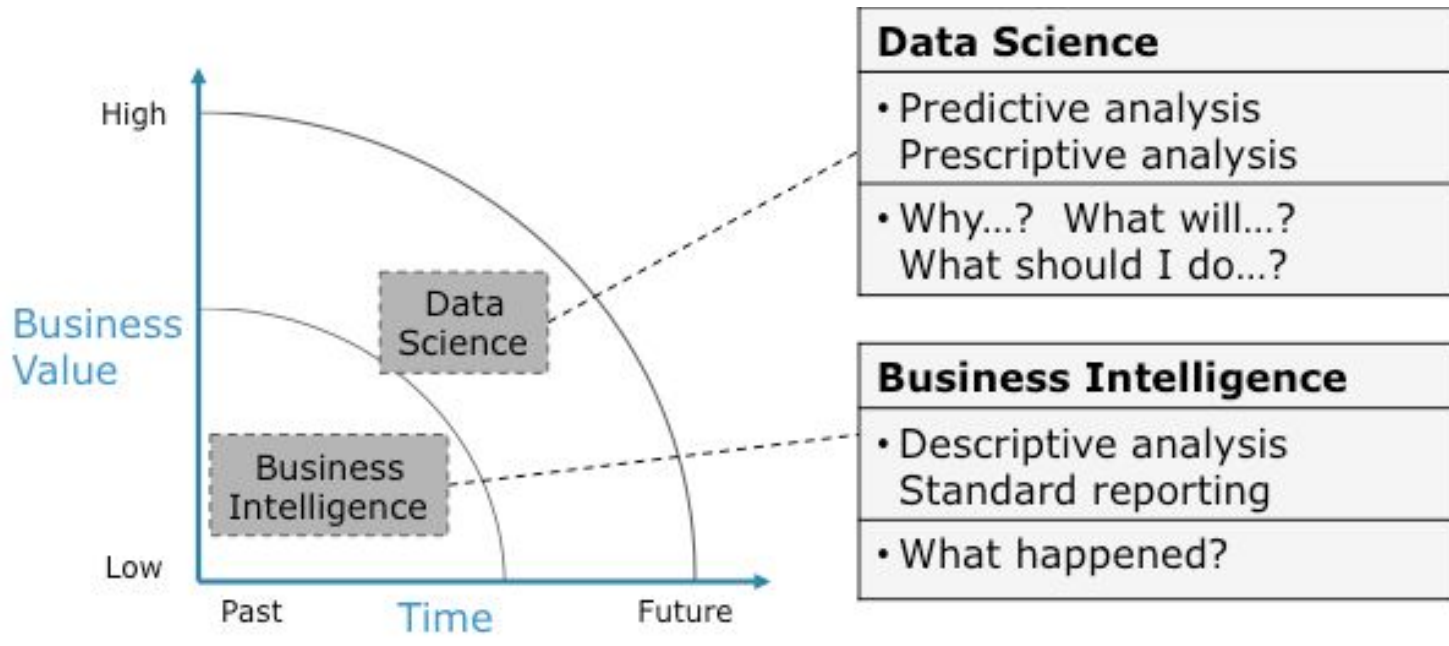


Qué es "Analytics"?

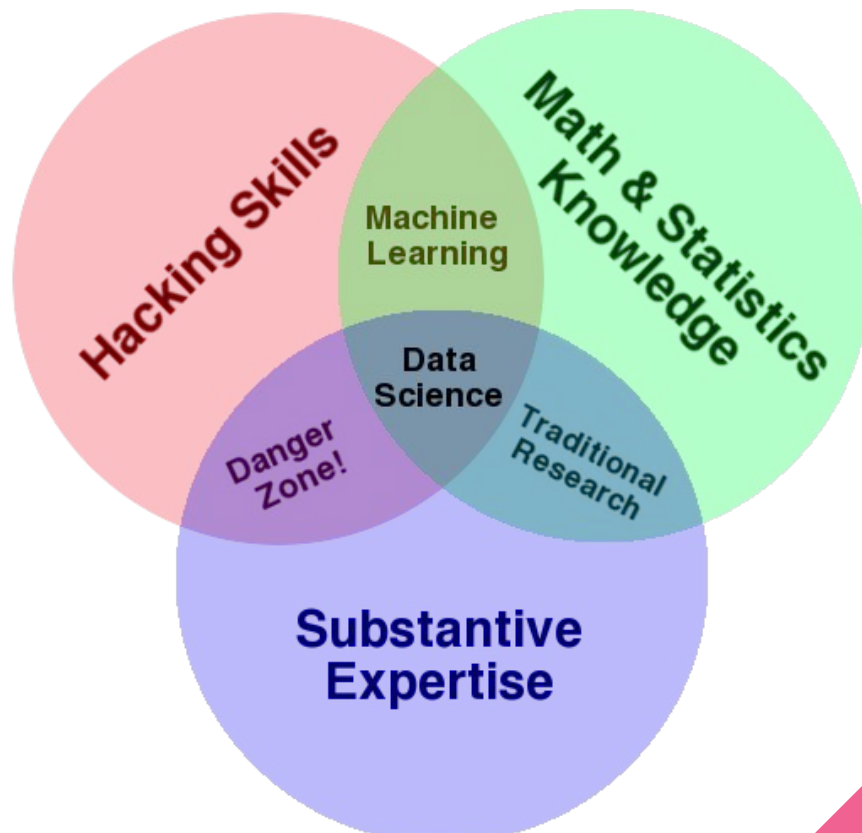


Qué es "Data Science"?

Qué distingue BI de Data Science?



Qué compone a Data Science?



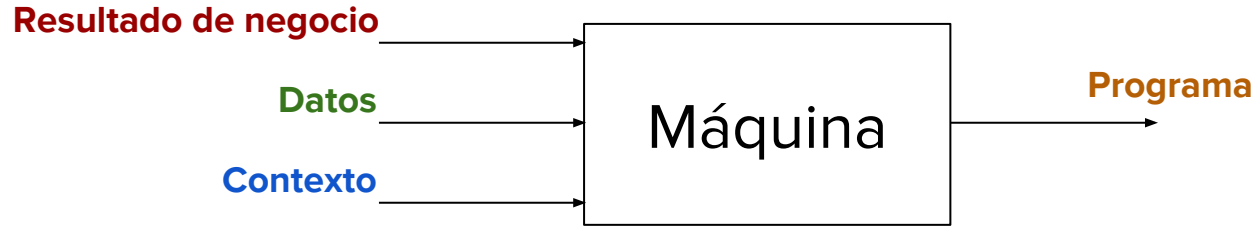
Qué es "Machine Learning"?

Qué es ML?

"Conjunción de Matemáticas, Estadística y Ciencias Computacionales para crear algoritmos de descubrimiento de funciones que partan el espacio de datos, ajusten a una curva sin conocer su origen, o detecten grupos emergentes."

- Dr. Fernando Esponda (ITAM)

Componentes de un proyecto de ML



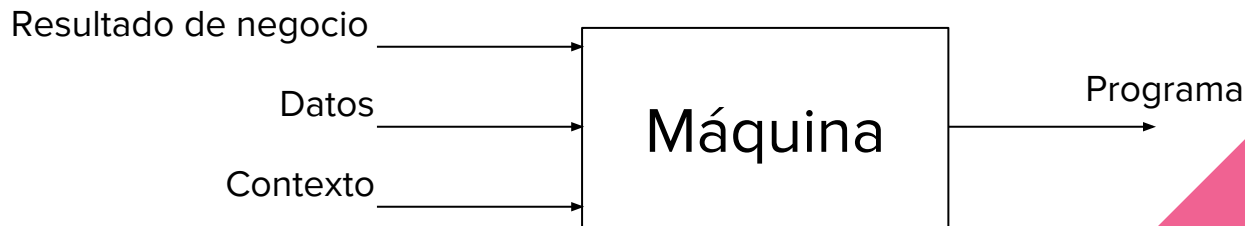
	Altura	Peso	Compleción	Talla	
<ul style="list-style-type: none">- Rangos- Medias- Máximo- Mínimo- Selección de obs					<ul style="list-style-type: none">- Patrones- Distribución- Clasificaciones- Inferencias- Ajuste curvas

Qué diferencia tiene con Desarrollo de SW?

Proyectos de Software



Proyectos de Machine Learning



Qué implica esta diferencia en los 2010s?

1960



**CEO NO
familiarizado con IT**

2010



**CEO familiarizado
con IT**

2010



**CEO NO
familiarizado con
DS**

2060?



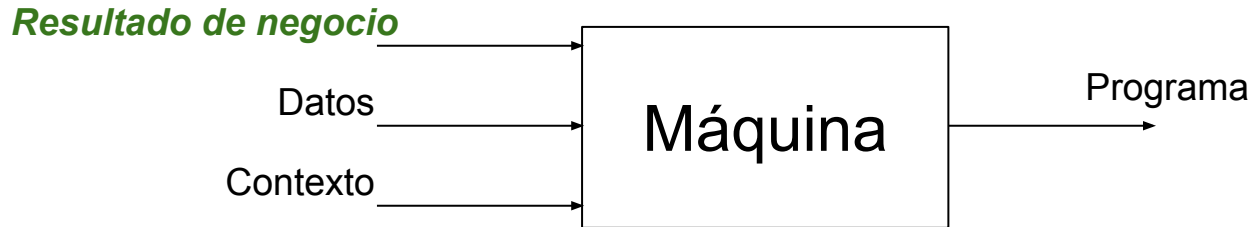
**CEO familiarizado
con DS**

Qué implica esta diferencia en los 2010s?

- CEO no sabe los detalles de los proyectos de DS / ML.
- Para él correlación == causalidad.
 - Y todo...TODO es Big Data.
 - Y todos los datos de la organización están limpios y disponibles.
 - Y si no, siempre está "análisis de sentimiento de los tuiters".
- Los proyectos de DS tienen fuerte componente tecnológico.
- Por tanto, los CEOs tratan proyectos de DS como si fueran de Desarrollo de Software.
- Los administran igual, les asignan los mismos recursos, las mismas fases, y los mismos KPIs.
- Este mismatch mete estrés en el proyecto, y, por presión del CEO, se degrada a lo más conocido, un proyecto de SW, dejando fuera a DS.

Tipos de Machine Learning

Supervisado



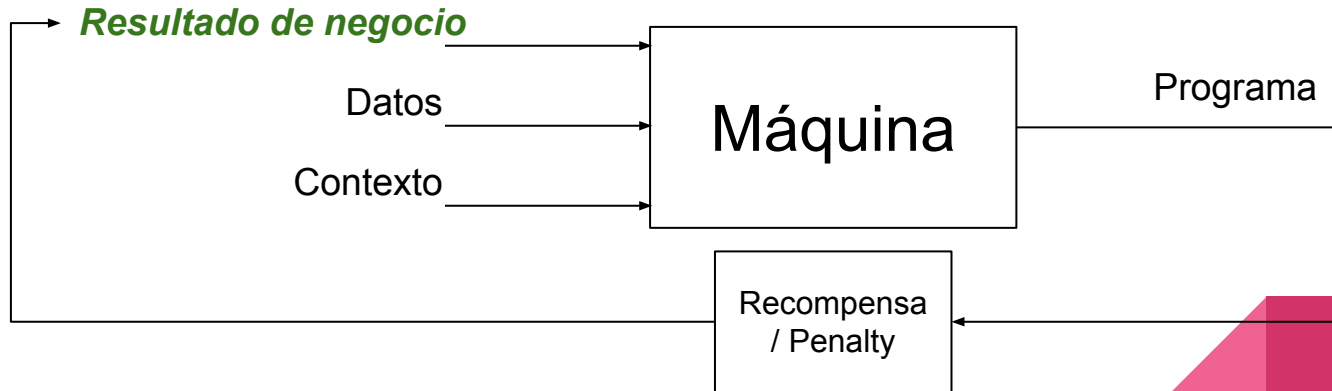
Tipos de Machine Learning

No-supervisado



Tipos de Machine Learning

Reinforcement



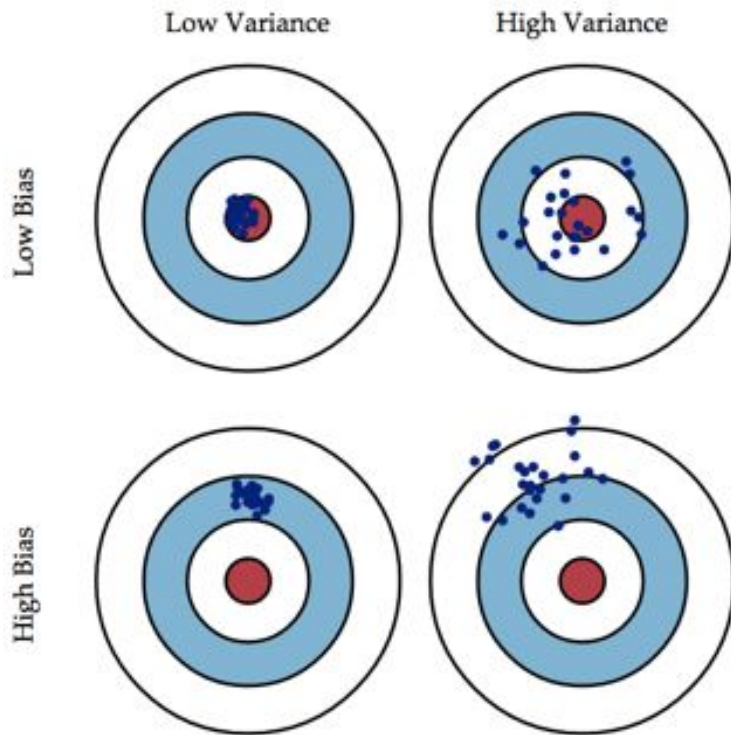
Objetivo del Machine Learning

Identificar
patrones
equivocándose
lo menor posible.



Cómo lo logra?

Balanceando
componentes del
error: sesgo +
varianza.



Ejemplo

Con datos de **17 mil 364 personas mayores de 18 años**, la **Cámara Nacional de la Industria del Vestido** encontró que **el hombre mexicano promedio pesa 74.8 kilos y mide 1.64 metros**, mientras que **las mujeres 1.58 metros de altura y 68.7 kilos de peso**.

Ejemplo

Posible sesgo de la muestra



Con datos de **17 mil 364 personas mayores de 18 años** la Cámara Nacional de la Industria del Vestido encontró que **el hombre mexicano promedio pesa 74.8 kilos y mide 1.64 metros**, mientras que **las mujeres 1.58 metros de altura y 68.7 kilos de peso**.

Ejemplo

Posible sesgo de la muestra

Con datos de **17 mil 364 personas mayores de 18 años**, la **Cámara Nacional de la Industria del Vestido** encontró que **el hombre mexicano promedio** pesa 74.8 kilos y mide 1.64 metros, mientras que **las mujeres** 1.58 metros de altura y 68.7 kilos de peso.

Sesgo de selección

Ejemplo

Posible sesgo de la muestra

Con datos de **17 mil 364 personas mayores de 18 años**, la **Cámara Nacional de la Industria del Vestido** encontró que **el hombre mexicano promedio** pesa **74.8 kilos** y mide **1.64 metros**, mientras que **las mujeres** **1.58 metros de altura** y **68.7 kilos de peso**.

Sesgo de selección

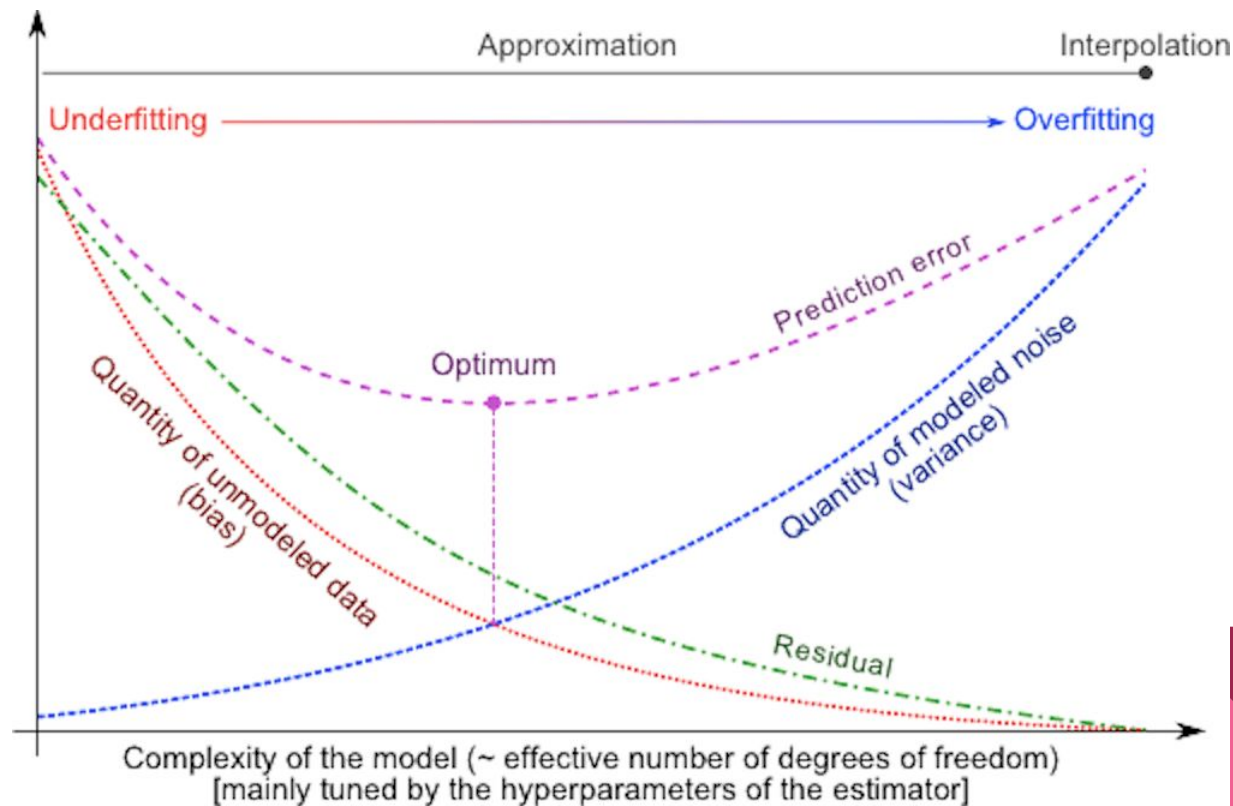
Varianza del fenómeno

Cómo balanceamos?

A mayor número de variables, mayor complejidad.

También mayor varianza.

Y menor sesgo.



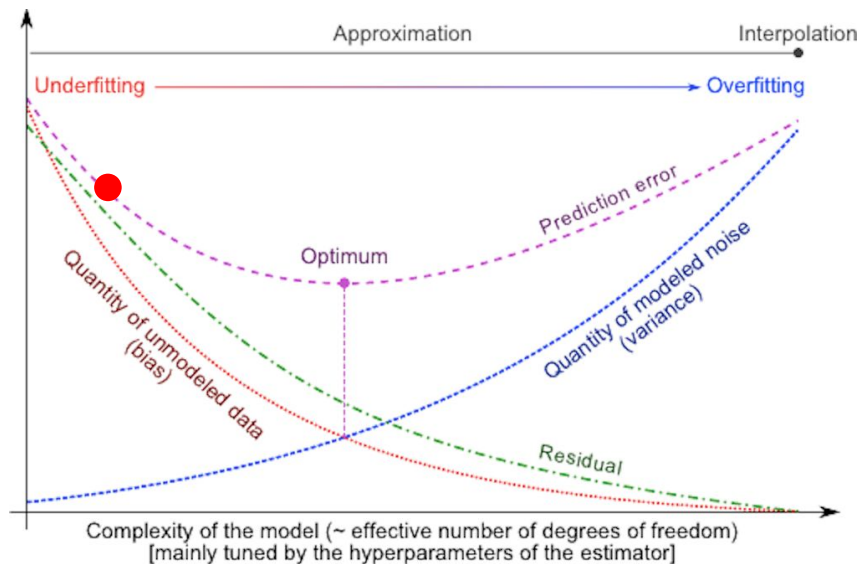
Ejemplo del Bias-Variance Tradeoff

X_n = Predictores. Y = Objetivo.

X_1	X_2

Y

Para predecir, por ejemplo, talla, necesitamos altura y peso.



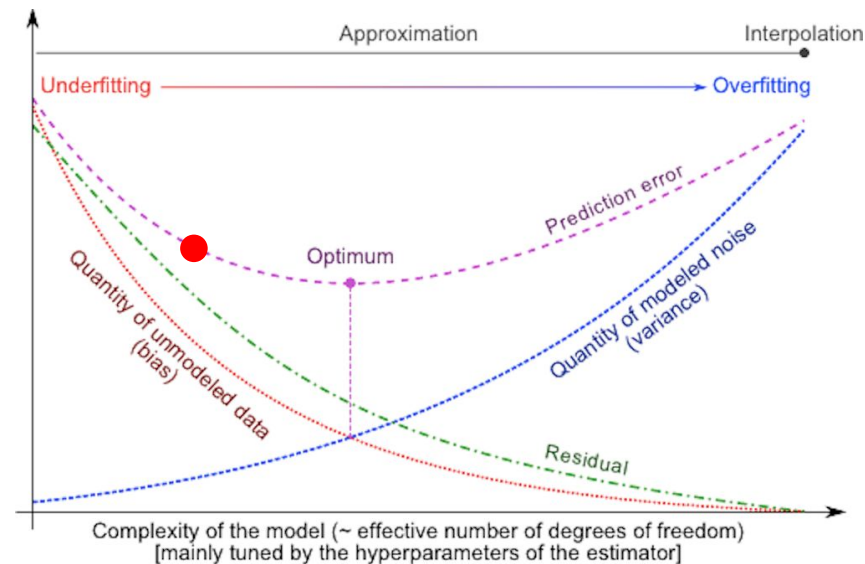
Ejemplo del Bias-Variance Tradeoff

X_n = Predictores. Y = Objetivo.

X1	X2	X3

Y

...y seguramente género.



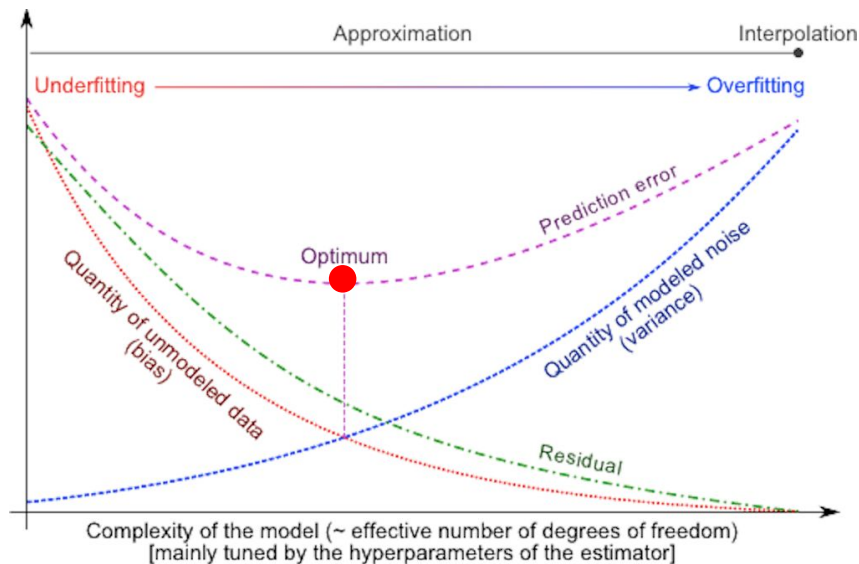
Ejemplo del Bias-Variance Tradeoff

X_n = Predictores. Y = Objetivo.

X1	X2	X3	X4

Y

Y ayudaría mucho tener quizá la variable categórica de "complexión".



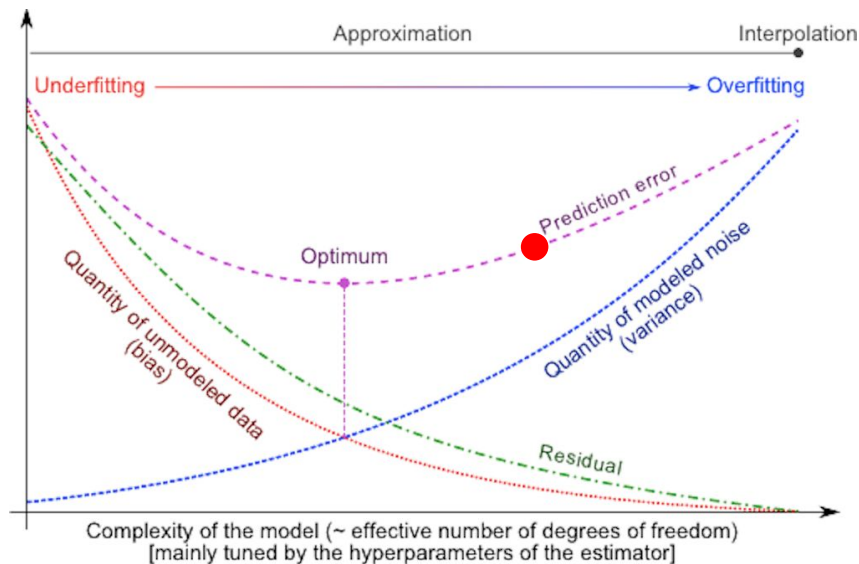
Ejemplo del Bias-Variance Tradeoff

X_n = Predictores. Y = Objetivo.

X1	X2	X3	X4	X5	X6

Y

Pero luego podemos emocionarnos y aventar variables que no aportan información, sino ruido, como el código postal o color de ojos.

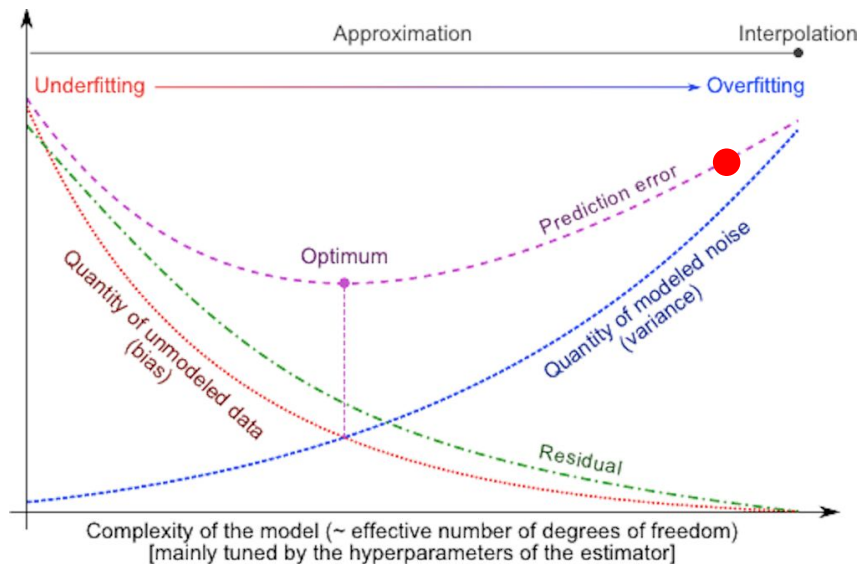


Ejemplo del Bias-Variance Tradeoff

X_n = Predictores. Y = Objetivo.

X1	X2	X3	X4	X5	X6	X7	X8	Y

O de plano el día de la semana, o el clima.
Este problema sucede frecuentemente
cuando quien está seleccionando variables
no tiene contacto con el negocio, ni sabe lo
que le duele.



Componentes del Error

$$\text{Error}_{\text{Total}} = \text{Error}_{\text{Varianza}} + \text{Error}_{\text{Sesgo}} + \text{Error}_{\text{Irreducible}}$$

Error total del modelo

Cómo reducimos?
Reduciendo los componentes

Error aportado por las varianzas de los predictores.

Cómo reducimos?
Agregando más observaciones.

Error aportado por el efecto de variables que no conocemos y por tanto no están modeladas.

Cómo reducimos?
Agregando mayor num de variables RELEVANTES.

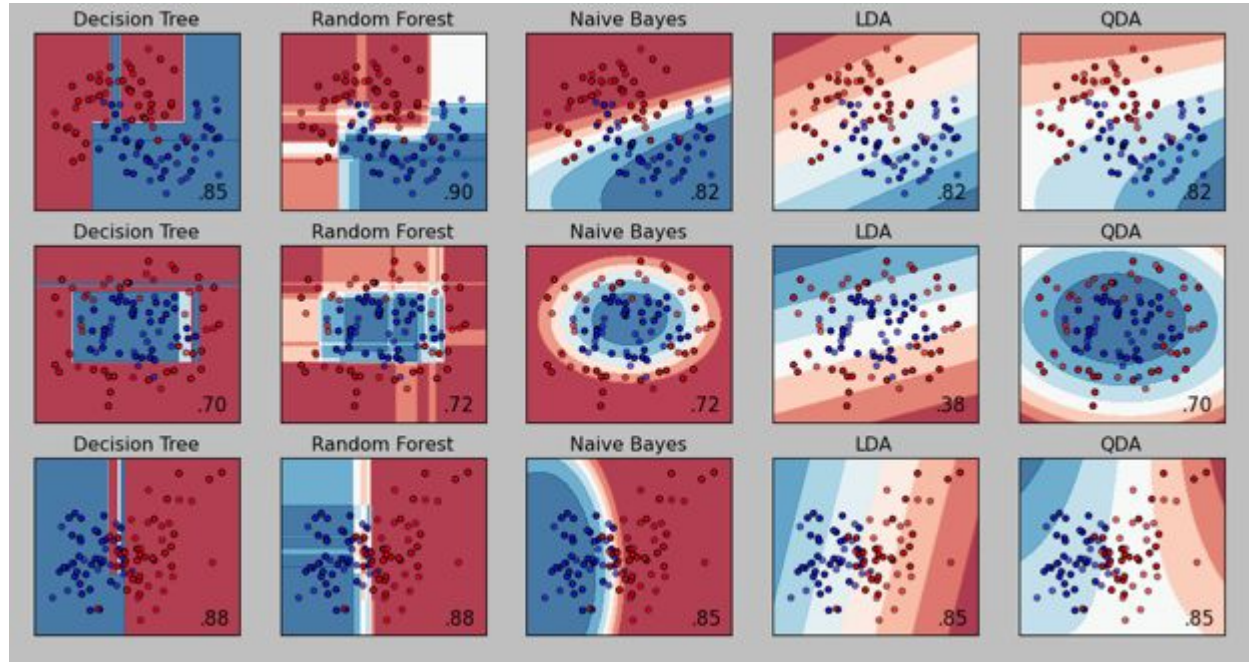
Ningún modelo será perfecto. Si lo es, probablemente sea overfitting. Este error representa la fracción de error que no podemos controlar sin romper el balance ya logrado.

Cómo reducimos? Posible, pero impráctico e inconveniente.

Usos del Machine Learning

Clasificación

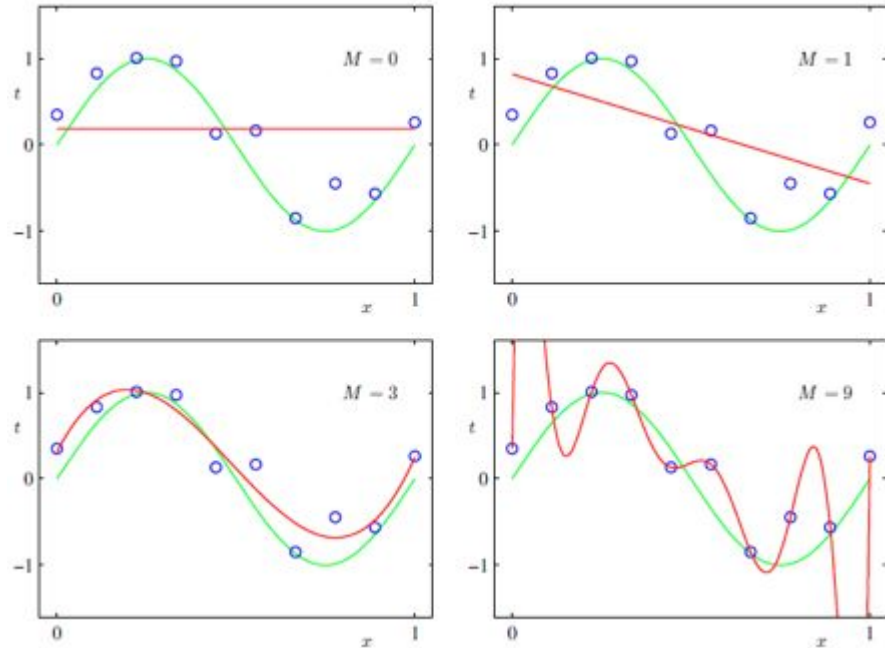
Partir el espacio de datos en N categorías con N-1 funciones trazadas en el espacio de datos.



Usos del Machine Learning

Regresión

Obtener 1 o varias funciones que ajusten a una curva continua existente en todo su espacio, o en fracciones del mismo.

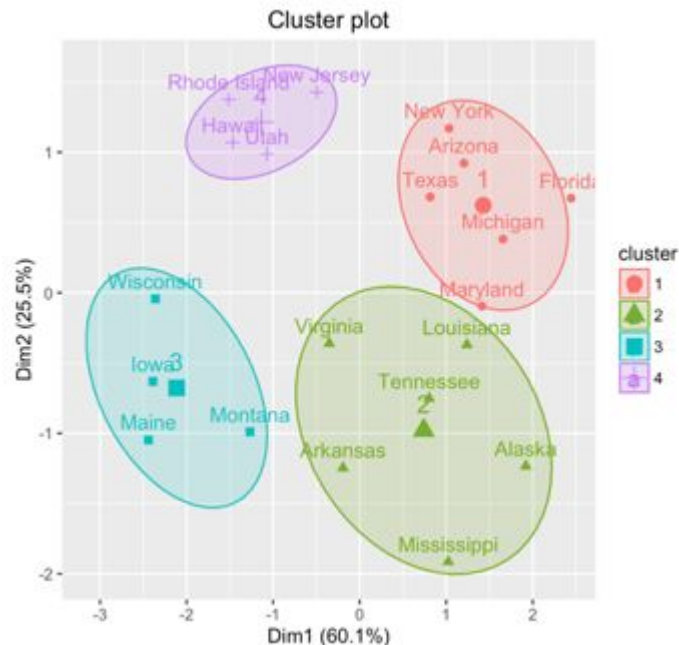


Usos del Machine Learning

Agrupamiento/Clustering

No supervisado!

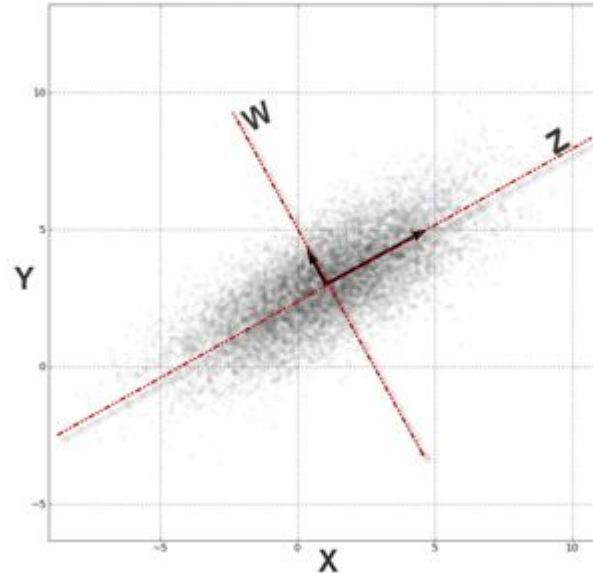
Descubrimiento de grupos emergentes estableciendo centroides y diferentes algoritmos de distancia.



Usos del Machine Learning

Reducción de dimensionalidad

Proyectar una representación del espacio de datos a un conjunto reducido de dimensiones y evaluar pérdida de información.



Y qué es Artificial Intelligence?

Qué es AI?

**Pensar como
humano**

**Pensar
racionalmente**

**Actuar como
humano**

**Actuar
racionalmente**

Qué es AI?

Pensar como humano

Pensar racionalmente

Actuar como humano

Actuar racionalmente

Parte VI

Use Cases de ML (buenos y malos)

- Buenos
 - Banca
 - Telco
 - Logística
 - Remesas
 - Medios
- Malos
 - Fashion
 - Internet
 - Telco
 - Twitter
 - Salud

Los Buenos

$B*n*m*x$

1. Conversión de cliente de nómina a TC en 29%.
2. \$2.7mmdp en revenue al año desde 2010.
3. Cómo lo hizo?
4. Clasificación!

UPS

1. Ahorro de combustible haciendo que camiones sólo den vuelta a la derecha.
2. Ahorro de \$47mdd al año.
3. Cómo lo hicieron?
4. Diseño de experimentos!

Famosa Telco Latinoamericana

1. Identificación de usrs consumiendo \$7K MXN semanales de tiempo aire en prepago.
2. Creación de producto de crédito de tiempo aire de hasta \$2K.
3. \$4mmdp al año de revenue.
4. Cómo lo hicieron?
5. Clustering!

Western Union

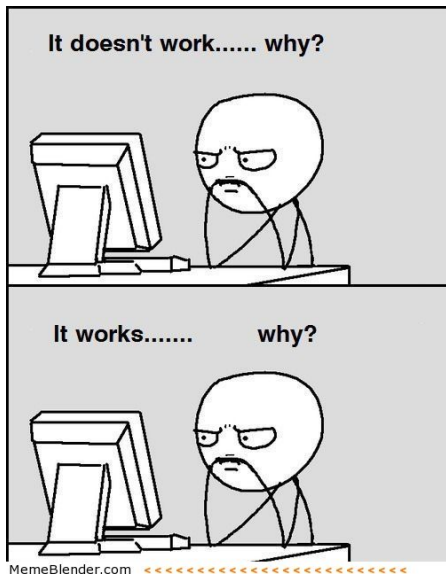
1. Prevención de fraude en remesas en automático y personalizado.
2. \$32mdd en ahorro operativo en 2012. \$21mdd son de transacciones detenidas al momento.
3. Cómo le hicieron?
4. Clasificación!
5. Similar a algoritmos de spam/ham.

Gr*p* *xp*ns*ón

1. Bajar bounce rate y mantener al visitante en sitios de las marcas del grupo.
2. Aumentar ad impressions.
3. Cómo lo están haciendo?
4. Recommender Systems!

Los Malos

Qué puede salir mal?



Telco importante dando créditos a sospechosos criminales.

Empresa importante de internet no le atina a predicción de AH1N1.

Gran empresa de software crea un bot sexualmente cargado y racista.

Empresa de internet clasifica foto de 2 afroamericanos como gorilas.

Crisis financiera de 2008.

Victoria de Trump.

Ejercicio

Diagnosticando problemas con proyectos de ML

Telco

Empresa detecta grupo de clientes que consumen \$7,000 en prepago a la semana con clustering.

Se crea un producto de crédito de hasta \$2,000 que se extiende cuando su crédito se acaba.

Dicho producto le representa \$13mdp al año a esta Telco.

Eventualmente se hace geocoding de sus direcciones.

Todos están en Sinaloa, Veracruz, Tamaulipas...ya se imaginarán de qué personas hablamos.

Qué falló?

Google Flu Trends

Google desarrolla un producto que pretende predecir outbreaks de AH1N1.

Se basa en núm. de búsquedas, geocoding, frecuencia de búsquedas, época del año.

Predice que en París habrá un outbreak de AH1N1. Francia pone en movimiento todo su aparato de protección civil. Al final no sucede nada, mas que el gasto innecesario de millones de Euros.

El modelo fue creado por programadores sin mucha conexión con la realidad.

Qué salió mal? Por qué se equivocó este modelo supervisado?

Microsoft Tay

Microsoft despliega Tay, un tweetbot que "aprende" del agregado de tweets del mundo angloparlante.

Tay, al cabo de 2 semanas de consumir el TL del mundo, se vuelve sexualmente cargado y racista.

Microsoft apaga Tay.

Qué sucedió? Cómo se pudo haber prevenido esto? Falló del todo?

Google Image Classifier

Google lanza su clasificador de imágenes para catalogado automático de fotos.

Un usuario sube la foto de 2 afroamericanos.

El Image Classifier los etiqueta como "gorilas".

Sucede una mini-catástrofe de PR para Google.

El clasificador lo hizo uno de los ingenieros de la empresa.

Cómo se pudo haber evitado este problema y seguir clasificando automáticamente?

Crisis Financiera de 2008

Se relajan criterios para préstamo hipotecario en EU. Gente que no podía acceder a créditos altos ahora puede hacerlo.

Dicha deuda "subprime" se empaqueta en productos estructurados llamados Collateral Debt Obligation, que ignoran de motu proprio las variables originales de los préstamos hipotecarios que dieron lugar la baja calificación que tienen.

Dichos CDOs se transactan como pan caliente en mercados no regulados (derivados OTS).

Cuando los préstamos hipotecarios escondidos en los CDOs caen en impago, éstos dejan de tener valor y se vuelcan sobre el mercado.

Cómo se pudo haber evitado que los CDOs se volvieran tóxicos?

Victoria de Trump

La empresa Cambridge Analytica crea grupos de FB "fantasma" cuyo único propósito es recoger perfiles y datos de sus afiliados.

Analizando dichos datos, se crean anuncios (a veces con contenido tergiversado) destinados a socavar el apoyo a Clinton y/o a aumentar el apoyo a Donald Trump.

La discusión política en FB se polariza.

Trump gana el voto colegiado.

Cómo pudimos haber evitado que Cambridge Analytica guiara de esa manera la opinión popular?

Sistema de Estimación de Reincidencia

Se desarrolla un algoritmo para clasificación de reincidencia delictiva para "minor offenders" y así solo encarcelar a sospechosos realmente peligrosos.

Se observa que se encuentra sesgado contra la población afroamericana.

Una mujer que robó una tienda para alimentar a su familia obtiene un score de reincidencia de 8.

Mientras que un asaltante blanco con historial de violencia recurrente con arma de fuego obtiene un score de 3.

Qué podemos hacer para evitar este error?

Clasificador de Profesores en Washington State

Consultora de Princeton "Mathematica Policy Research" hace un modelo (no revela sus detalles) para calificar a profesores de primaria de escuelas públicas en 2010-2011.

206 maestros son despedidos sin explicación. Sarah Wyzocki, una de ellas, comienza a investigar con ayuda del Washington Post, y pedir accountability.

Descubren que unas de las principales variables del modelo son los scores en mates y lectura de sus alumnos en pruebas estándar del estado. Los alumnos son de zonas problemáticas. Descubren que el modelo no considera situaciones de pobreza, violencia intrafamiliar ni cuestiones nutricionales de los niños.

Una vez publicada la investigación, se descubren consistentemente en los school districts mejor evaluados exámenes con borrones y correcciones.

Cuál fue el problema con este modelo? Cómo pudimos haber evitado que Sarah Wyzocki fuera despedida?

Parte VIII

Machine Ethics

Algoritmos que discriminan

Transparencia Algorítmica

Implicaciones para el ML

Deficiencias del DL

Marco ético de la ACM

"The Modeller's Manifesto"



FACT:



Los algoritmos discriminan*.

**Discriminar*

Del lat. *discrimināre*. tr. Seleccionar excluyendo.

Algoritmo para asignar fianzas y predecir reincidencia

Two Petty Theft Arrests

	
VERNON PRATER	BRISHA BORDEN
LOW RISK 3	HIGH RISK 8

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

Algoritmo para
asignar fianzas y
predecir reincidencia

Sesgado VS
afroamericanos* :/

Two Petty Theft Arrests

 <p>VERNON PRATER</p> <p>Prior Offenses 2 armed robberies, 1 attempted armed robbery</p> <p>Subsequent Offenses 1 grand theft</p> <p>LOW RISK 3</p>	 <p>BRISHA BORDEN</p> <p>Prior Offenses 4 juvenile misdemeanors</p> <p>Subsequent Offenses None</p> <p>HIGH RISK 8</p>
--	--

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

* <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Búsqueda de "CEO" en Google Images

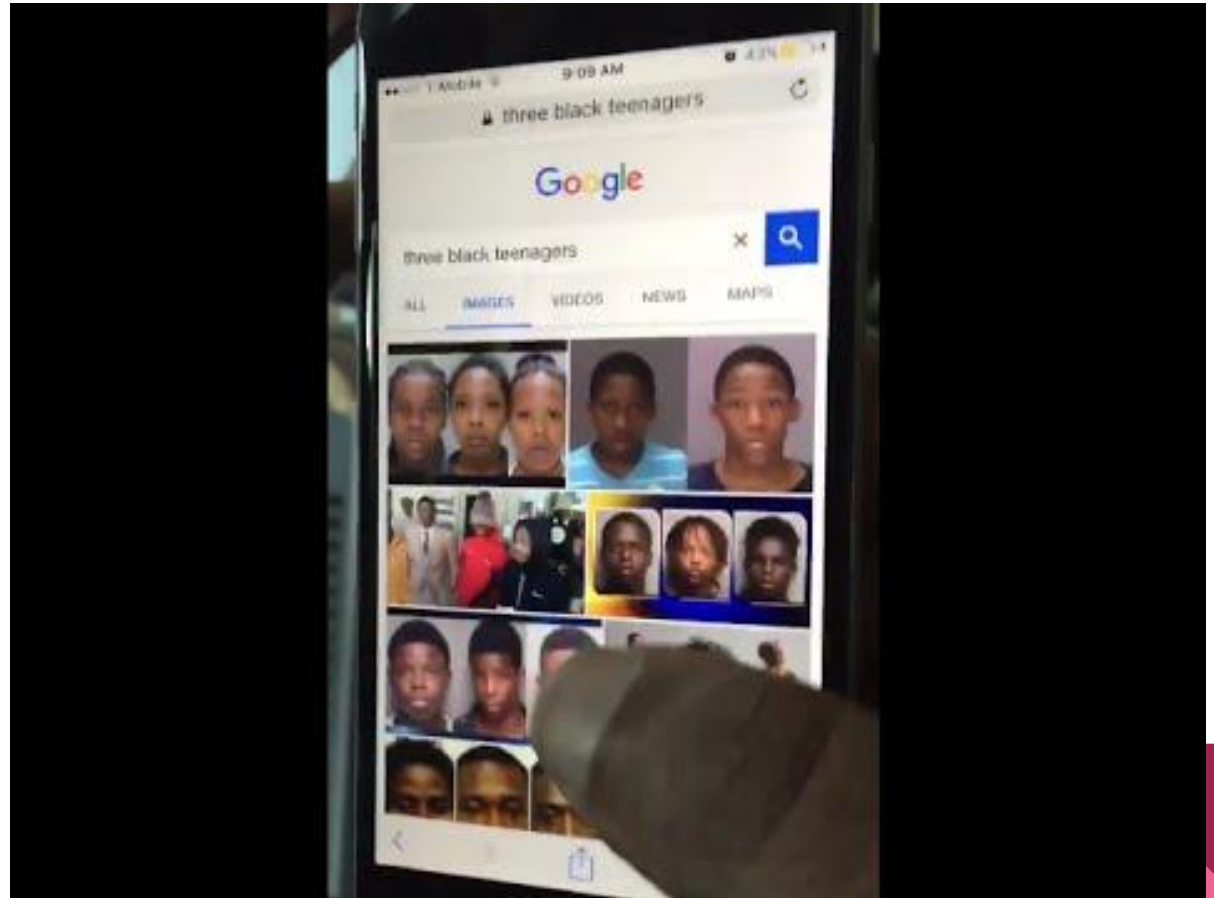


Mujeres en los resultados = 11/100

Mujeres CEO en EU = 27/100*

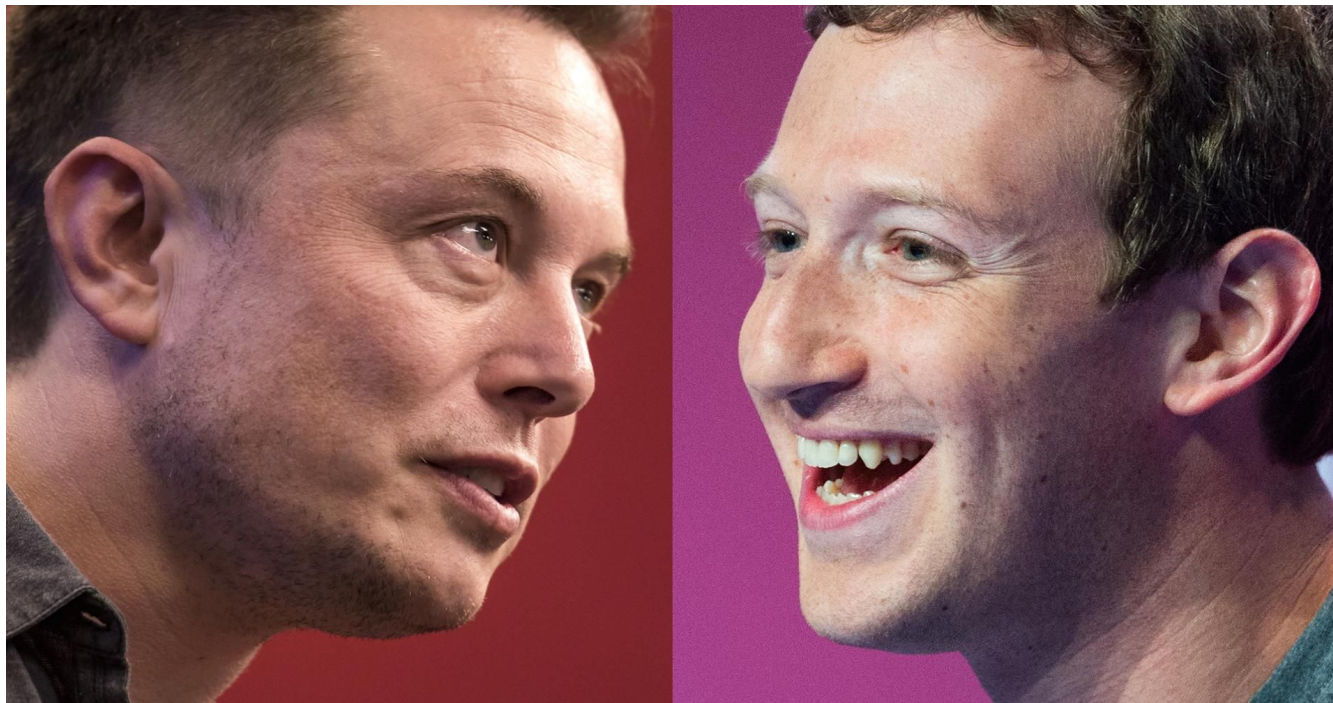
* https://www.eurekalert.org/pub_releases/2015-04/uow-wac040915.php

"Three black teenagers" VS
"Three white teenagers"*



* <https://www.theguardian.com/technology/2016/jun/09/three-black-teenagers-anger-as-google-image-search-shows-police-mugshots>

Musk VS Zuckerberg



Elecciones 2018



Elecciones 2018



Causas

1. Falta de contexto

Equipo dejó fuera al subject matter expert.

2. Correlación = Causalidad

Equipo dejó fuera a matemáticos/físicos.

3. Ninguna de las anteriores

Todo se hizo bien, los datos están sesgados de origen.

Causas

1. Falta de contexto


Equipo dejó fuera al subject matter expert.

2. Correlación = Causalidad

Equipo dejó fuera a matemáticos/físicos.

3. Ninguna de las anteriores

Todo se hizo bien, los datos están sesgados de origen.



"A big step towards countering discriminatory algorithms is the ability to understand them..."

- Ethan Chiel, writer @ Fusion

Abril 2016: Unión Europea*

European Union regulations on algorithmic decision-making and a “right to explanation”

Bryce Goodman,^{1*} Seth Flaxman,²

¹Oxford Internet Institute, Oxford

1 St Giles', Oxford OX1 3LB, United Kingdom

²Department of Statistics, University of Oxford,
24-29 St Giles', Oxford OX1 3LB, United Kingdom

* <https://arxiv.org/pdf/1606.08813.pdf>

Qué significa esto para la
sociedad?

No te aceptaron en la París IV?

No te dieron tu crédito hipotecario?

Te recomendaron un tratamiento
médico?

Fuiste seleccionado para programas
de gobierno en un tier X?

No fuiste aceptado en el Eurovision?

**Tienes derecho a
saber por qué!**

**Tienes derecho a
una explicación!**

Qué significa esto para DS y ML?

Antes:

Buena Predicción > Interpretabilidad

Ahora:

Interpretabilidad >> Buena Predicción

Qué significa esto para el ML supervisado?

1. Privilegiar simplicidad.
2. Dedicar tiempo al Feature Selection.
3. Para tener modelos con menos variables.
4. Y poder entrenar varios en un tiempo menor.
5. Y así evitar la maldición de la dimensionalidad.
6. Y finalmente reducir el overfitting.

Qué significa esto para el ML no supervisado?

1. Privilegiar la reproducibilidad*
2. Elegir algoritmos parametrizables (DBSCAN > K-means).
3. Establecer procesos formales de diseño de experimentos.

* <https://www.forbes.com/sites/quora/2017/02/09/how-the-reproducibility-crisis-in-academia-is-affecting-scientific-research/#1aa4d3853dad>

Qué significa esto para la Ingeniería de Datos?

1. Incorporar Github al proceso de DS.
2. Levantar infras para entrenamientos en paralelo.
3. Levantar pipelines de reentrenamiento.
4. "Solo los salvajes sobrescriben datos".
5. Envolver la DS en Ingeniería de Software.



Y qué significa para
el Deep Learning?

Qué significa esto para el Deep Learning?

Explicar decisiones en términos de weights no es admisible, ni accionable para el afectado*

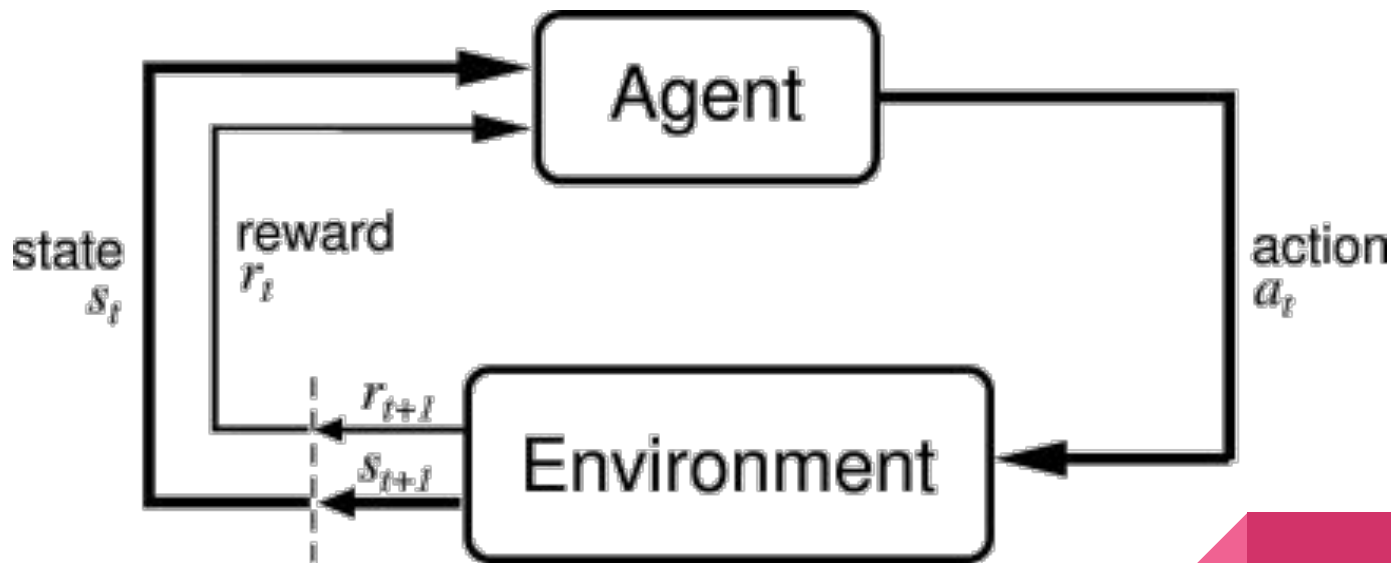
* <https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/>

Entonces ya no lo usamos?

Cuando la distancia entre una decisión, y su recompensa / penalización sea directa, usa DL*

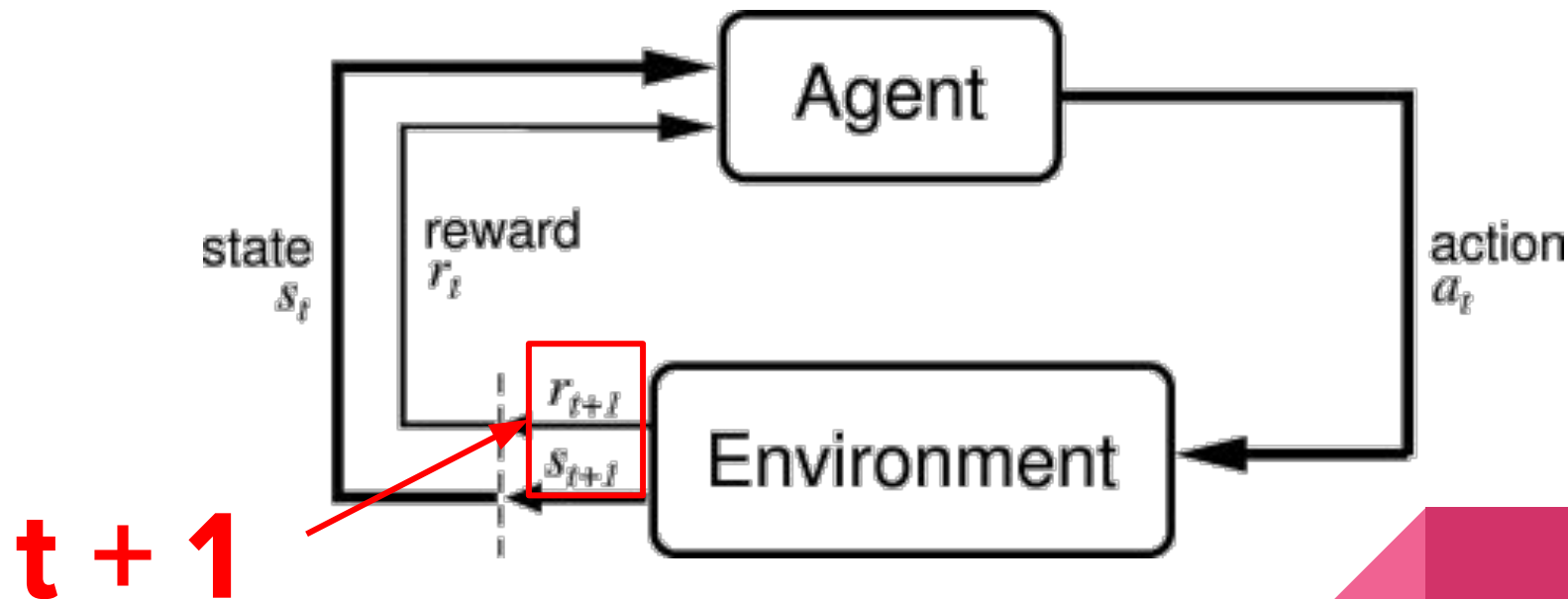
* <http://www.wired.co.uk/article/google-ai-montezuma-revenge>

Relación entre Reward y Penalización



* <http://www.wired.co.uk/article/google-ai-montezuma-revenge>

Relación entre Reward y Penalización



$t + 1$

* <http://www.wired.co.uk/article/google-ai-montezuma-revenge>

Y cuando no lo sea?

DARPA¹, MIT², Cambridge, Oxford, DATANK, et al ya están en ello.

1. <http://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf>
2. <http://news.mit.edu/2016/making-computers-explain-themselves-machine-learning-1028>

Y mientras?

Deep Forest: Towards An Alternative to Deep Neural Networks

Zhi-Hua Zhou and **Ji Feng**

National Key Laboratory for Novel Software Technology

Nanjing University, Nanjing 210023, China

{zhouzh, fengj}@lamda.nju.edu.cn

* <https://arxiv.org/pdf/1702.08835.pdf>



Hay un marco ético
para esto?

ACM*

Enero 2017

Awareness of Bias

Access & Redress

Accountability

Explanation

Data Provenance

Auditability

Validation & Testing

*

https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf

The modeller's oath (Wilmott & Derman, 2008)

- 1. I will remember that I didn't make the world, and it doesn't satisfy my equations.**
- 2. Though I will use models boldly to estimate value, I will not be overly impressed by mathematics.**
- 3. I will never sacrifice reality for elegance without explaining why I have done so.**
- 4. Nor will I give the people who use my model false comfort about its accuracy. Instead, I will make explicit its assumptions and oversights.**
- 5. I understand that my work may have enormous effects on society and the economy, many of them beyond my comprehension.**